PROJECT PHASE 2 REPORT

ON

Concatenation of Attention Enhanced Spatial and Temporal Features for Violence Detection from Videos

Submitted by

Ajay T Shaju (SJC20AD004)

Emil Saj Abraham (SJC20AD028)

Justin Thomas Jo (SJC20AD046)

Vishnuprasad K G (SJC20AD063)

to

the APJ Abdul Kalam Technological University

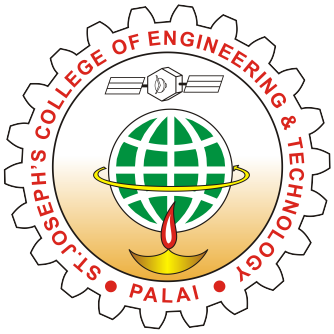
in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in

Artificial Intelligence and Data Science



Department of Artificial Intelligence and Data Science
St. Joseph's College of Engineering and Technology, Palai

May : 2024

Declaration

We undersigned hereby declare that the project phase 2 report on **"Concatenation of Attention Enhanced Spatial and Temporal Features for Violence Detection from Videos"**, submitted for partial fulfillment of the requirements for the award of the degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala, is a bonafide work done by us under the supervision of **Dr. Renjith Thomas**. This submission represents our ideas in our own words and where ideas or words of others have been included. We have adequately and accurately cited and referenced the original sources. We also declare that we have adhered to the ethics of academic honesty and integrity and have not misrepresented or fabricated any data, idea, fact, or source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma, or similar title of any other University.

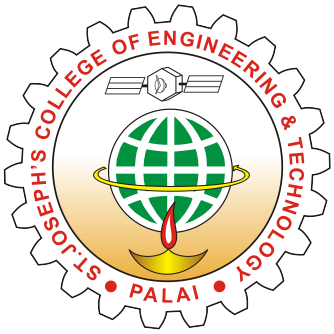
Name and Signature of Students

Ajay T Shaju (SJC20AD004)
Emil Saj Abraham (SJC20AD028)
Justin Thomas Jo (SJC20AD046)
Vishnuprasad K G (SJC20AD063)

Place: Choondacherry
Date: 03-05-2024

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY, PALAI

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE



CERTIFICATE

This is to certify that the report entitled "**Concatenation of Attention Enhanced Spatial and Temporal Features for Violence Detection from Videos**" submitted by **Ajay T Shaju (SJC20AD004), Emil Saj Abraham (SJC20AD028), Justin Thomas Jo (SJC20AD046), and Vishnuprasad KG (SJC20AD063)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Artificial Intelligence and Data Science is a bonafide record of the project work carried out by them under my guidance and supervision.

Project Guide
Dr.Renjith Thomas
Head of the Department
Department of AD

Project Coordinator
Mr.Jacob Thomas
Assistant Professor
Department of AD

Head of Department
Dr.Renjith Thomas
Associate Professor
Department of AD

Place: Choondacherry
Date: 03-05-2024

Acknowledgement

The success and final outcome of this project phase 2 required a lot of guidance and assistance from many people, and we are extremely privileged to have received their support throughout the completion of this project. We would like to express our respect and gratitude to the management of St. Joseph's College of Engineering and Technology for providing us with the opportunity and platform to work on this project. A special word of thanks goes to our beloved Principal, **Dr. V. P. Devassia**, for providing invaluable support and necessary facilities to carry out this project. We are extremely indebted to our project guide **Dr. Renjith Thomas**, Associate Professor & Head of the Department of Artificial Intelligence and Data Science, for his valuable suggestions and encouragement throughout the course of this project. His keen interest and guidance in this project provided us with all the necessary information for developing a robust system. We would also like to express our gratitude to our project coordinator, **Mr. Jacob Thomas**, Assistant Professor in the Department of Artificial Intelligence and Data Science, for his valuable suggestions and guidelines during the entire duration of this project. We truly appreciate his contributions and technical support in preparing this report.

We also extend our thanks to Dr.Deepa V, Assistant Professor, School of AI and Robotics, MG University, Kottayam, Mr. Milton John, Software Engineer, Amazon Bengaluru, and Mr. Sandeep C George, Senior Safety Officer, ThyssenKrupp Abu Dhabi for their amazing thoughts, help, and recommendations. Also, we thank all our friends, and teaching and non-teaching staff members of the Department of Artificial Intelligence and Data Science who have offered unparalleled help in needful times. All of these people have played a crucial role in the successful completion of our project phase 2 work.

Ajay T Shaju

Emil Saj Abraham

Justin Thomas Jo

Vishnuprasad KG

Abstract

The escalation of violence and disturbances in public areas like the stampede on the CUSAT Campus of Kerala, necessitates advancements in surveillance technology to ensure safety and security. The project titled ***"Concatenation of Attention Enhanced Spatial and Temporal Features for Violence Detection from Videos"*** addresses this need by developing a deep learning model that effectively combines spatial and temporal information for accurate violence detection in video feeds. Utilizing the MobileNetV2 architecture as a feature extractor Convolutional Neural Network (CNN), enhanced with a spatial attention mechanism, the model focuses on important spatial features in video frames to identify violent actions. To capture the movement and changes over time, which are necessary for detecting violence, the model integrates a Long-Short Term Memory (LSTM) network that processes information from both global average and max pooling layers. This approach not only improves the detection accuracy but also helps the model adapt to different scenarios and camera angles through data augmentation techniques. The training process is carefully managed with strategies such as early stopping, adaptive learning rate adjustments, and model checkpointing to optimize performance, prevent overfitting, and enable retraining. Thus embracing innovative AI approaches to provide lightweight, fast, and efficient model for monitoring systems with the capability to detect violent behaviors quickly and reliably, thereby contributing to a safer Earth for everyone.

List of Abbreviations

2D 2-dimensional

3D 3-dimensional

AI Artificial Intelligence

AOAV Action on Armed Violence

AUC Area Under Curve

BiLSTM Bidirectional Long Short Term Memory

C3D Convolutional 3D

CAM Class Activation Map

CAST Concatenation of Attention enhanced Spatial and Temporal features

CCTV Closed-Circuit Television

CNN Convolutional Neural Network

ConvLSTM Convolutional Long Short Term Memory

CSV Comma Separated File

CUSAT Cochin University of Science and Technology

DL Deep Learning

FN False Negative

FP False Positive

GPU Graphics Processing Unit

ICDICI International Conference on Data Intelligence and Cognitive Informatics

IEEE Institute of Electrical and Electronics Engineers

IoT Internet of Things

LaM-2SRN Local Features Enhanced and Moving target detected 2Stream-ResNet

LSTM Long Short Term Memory

ML Machine Learning

OOM Out-Of-Memory

RAM Random Access Memory

ResNet Residual Network

RGB Red Green Blue

RNN Recurrant Neural Network

ROM Read Only Memory

RWF Real-World Fights

SepConvLSTM Separable Convolutional LSTM

SPIL Skeleton Points Interaction Learning

SSHA Semi-Supervised Hard Attention

TN True Negative

TP True Positive

UCF University of Central Florida

VRAM Video RAM

XAI Explainable AI

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In today's interconnected and increasingly complex world, safeguarding public safety has become critical to governments, law enforcement agencies, and communities globally. In the middle of many challenges faced by modern societies, the rise in violent incidents presents a particularly pressing threat, with outbreaks of unrest and disorder occurring all too frequently. Recognizing the urgency of addressing this issue, this work advocates for the integration of AI and deep learning technologies into public safety frameworks [1]. The percentage change in peace score from the year 2008 to 2023 is shown below in Figure 1.1.

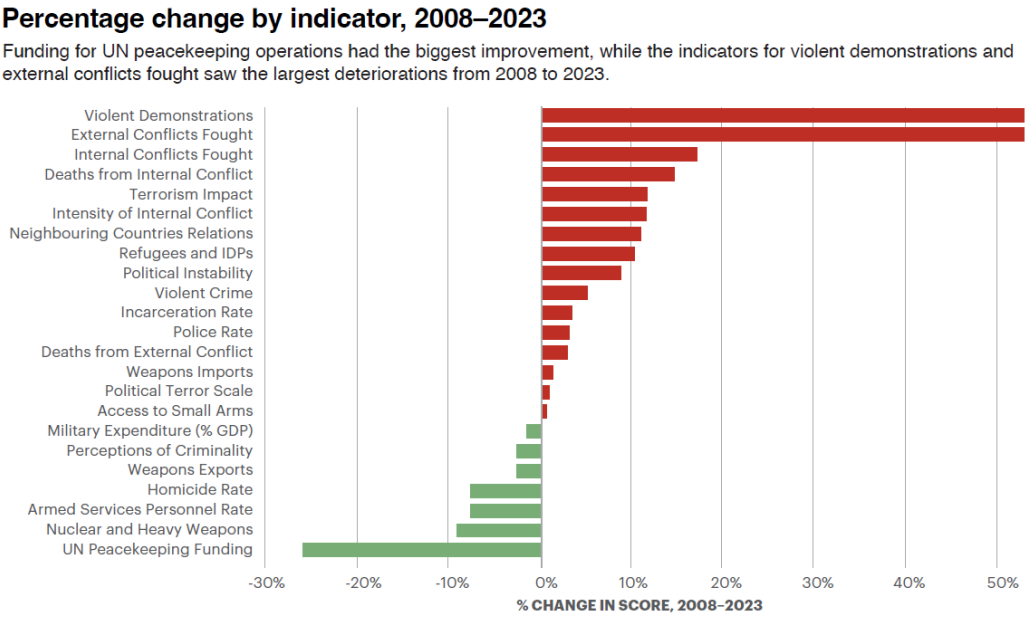


Figure 1.1: Percentage Change by Indicator, 2008–2023

Figure 1.1 shows that while there's been a positive development with a 50% increase in funding for UN peacekeeping, a closer look reveals a concerning trend. Since 2008, the world has grown less peaceful. Indicators surrounding conflict paint an unapproachable picture: a rise in violent demonstrations, an increase in the number of external and internal conflicts fought, and a surge in deaths caused by external conflicts. Furthermore, the number of refugees and instances of political instability have grown significantly. However, there are a few flashes of hope. Relations between neighboring countries have shown slight improvement, and there have been small decreases in both police rates and prison rates. Overall, the data suggests a world in need of solutions to promote peace and stability.

By harnessing the analytical skills of AI algorithms and the predictive capabilities of deep learning models, the team proposes a transformative approach to the detection and prevention of violence. The methodology involves the aggregation and analysis of data from diverse sources, including surveillance systems, social media platforms, and sensor networks. Through sophisticated data processing and pattern recognition techniques, the framework aims to identify early warning signs of potential disturbances, enabling proactive intervention by law enforcement agencies and security personnel. The strategic deployment of AI-driven insights and intervention strategies holds the promise of empowering authorities to anticipate and defuse volatile situations before they escalate into violence. By leveraging advanced predictive analytics and situational awareness tools, the framework seeks to enhance the effectiveness and efficiency of public safety efforts, thereby minimizing the risk of harm to individuals and communities.

Moreover, this work underscores the importance of considering ethical considerations and societal implications in the adoption of AI for violence detection and prevention [2]. As with any technology, the responsible use of AI requires careful attention to issues such as privacy, bias, and accountability. By addressing these concerns and encouraging transparency and accountability in AI-driven decision-making processes, therefore it ensures that efforts to enhance public safety are aligned with fundamental principles of justice and human rights.

The integration of cutting-edge technologies, particularly AI, holds the key to reshaping public safety and security. Embracing these advancements and promoting collaboration among diverse stakeholders, including government bodies, law enforcement agencies, technologists, and communities, is essential. By prioritizing inclusivity and equity in AI development it ensures the needs of all members of society, paving the way for safer, more resilient communities characterized by trust, transparency, and vibrant social interaction.

1.1 Background and Motivation

Violence, in its various forms, presents significant threats to public safety, organizational security, and individual well-being. Incidents of violence occur in diverse settings, including public spaces, workplaces, and homes, resulting in severe consequences such as injuries, fatalities, and infrastructure damage. Violence not only harms its direct victims but also spreads a broad impact on society. The trend in the global economic impact of violence through the years 2008 to 2022 is shown below in Figure 1.2

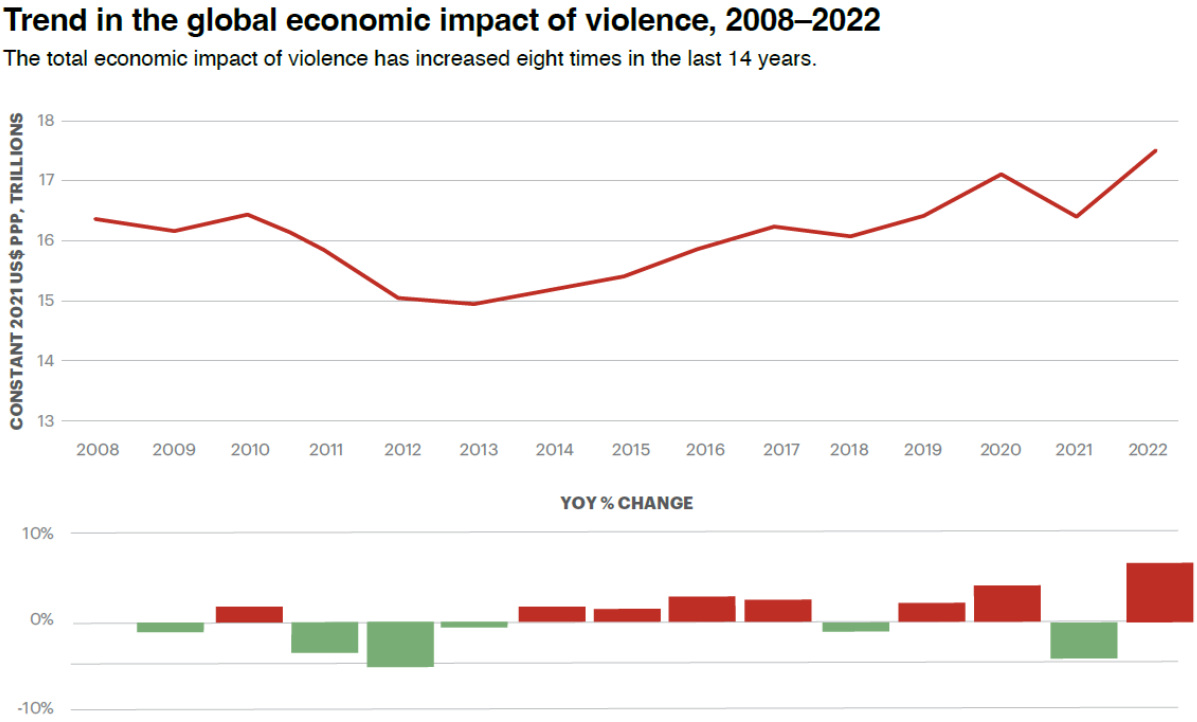


Figure 1.2: Trend in the Global Economic Impact of Violence, 2008–2022

On analyzing Figure 1.2, in 2008 the economic impact of violence was estimated to be around $10 trillion (in constant 2021 US dollars). By 2022, the economic impact of violence had increased to around $18 trillion. This represents an increase of 80% over the period. The graph also shows the economic impact of violence has increased significantly, such as in 2017(10%) and the economic impact of violence has decreased, in 2012(by 10%).

The primary effect of violence is that it breaks down trust and unity among communities. When individuals feel unsafe in their neighborhoods or workplaces, social interactions become strained, and community bonds weaken. Fear of violence can lead to social isolation and increase feelings of loneliness.

Additionally, violent incidents impose substantial economic burdens on society, involving costs associated with medical treatment, criminal justice proceedings, and victim rehabilitation services. Violent activities are not only between humans, humans to animals, and animals to humans, but human to materials is also possible. For example, if a stressed person breaks down a pipeline in a factory can cause injuries to others as well. Hence, it brings the need to prevent violent incidents as soon as possible before they go to extreme conditions.

Violent incidents among the public can be identified through surveillance cameras [3]. Despite advancements in video analysis, many surveillance systems still rely on human operators to monitor live feeds or review recorded footage. This process can be labor-intensive and prone to errors, as operators may miss violent incidents due to factors like fatigue, distractions, or the sheer volume of footage to review.

Even if a violent incident is captured on camera, the effectiveness of the response depends on how quickly it's detected and acted upon. Manual monitoring may not always ensure timely intervention, allowing incidents to escalate before appropriate measures are taken [4]. As a result, it is very essential to develop a system that can leverage cutting-edge technologies to enhance the detection and timely alerting for the prevention of violence. This can be achieved through leveraging the power of AI, machine learning, and computer vision techniques which enable the system to analyze live video feeds or recorded footage to automatically identify violent behaviors or patterns associated with the video.

Motivation

The motivation behind the development of a violence detection system is followed by real-life incidents that highlight the urgent need for improved methods for early analyzing and alerting the possible violent incidents that can happen in public spaces, which helps the relevant authorities take action to prevent them. One such incident that exemplifies this need is the tragic stampede that occurred during a cultural festival at Cochin University of Science and Technology (CUSAT) on 25th November 2023. A news clip of the same incident is shown in Figure 1.3

Figure 1.3: NEWS Clip of the CUSAT Stampede Incident, November 25, 2023, by India Today NEWS Channel

The CUSAT stampede, which resulted in multiple fatalities and injuries, was a reminder of the potential consequences of overcrowding, mismanagement, and the rapid escalation of violence in crowded environments. What began as a festive celebration quickly overturn into chaos and panic, as attendees struggled to navigate overcrowded pathways and exits, leading to a stampede that claimed the lives of innocent individuals.

This tragic event not only reminds the limitations of existing security measures but also highlights the need for more advanced technologies AI, ML, etc in detecting and mitigating violence in real time. Traditional methods were mostly reactive, meaning they would respond after something bad had already happened [4]. This delay made it harder to prevent violence from getting worse. Also, modern cities are so complex, with lots of people and different cultures mixing, which makes it even tougher for authorities to keep an eye on things. That's where AI steps in. It's like having a super-smart detective that can analyze tons of different data in real-time, like security camera footage and sensor readings [1]. When AI detects unusual behavior or signs of trouble early, it can alert the relevant authorities, enabling them to step in and prevent the situation from escalating.

Action on Armed Violence (AOAV) records, investigates, and disseminates evidence of armed violence against civilians worldwide, to ensure the respect and protection of their rights and to end armed violence against civilians in conflict. Some of the records of information are mentioned below in Table 1.1 and Table 1.2 and Figure 1.4:

Table 1.1: Annual Casualties Caused by Violence

Year	Total
2022	6,886
2023	15,305

The key findings from the global explosive violence monitor report by Action on Armed Violence underscore a concerning escalation in civilian casualties and incidents of explosive weapon use worldwide in 2023. The significant increase in civilian fatalities, a rise in explosive weapon use, and the prevalence of air-launched attacks highlight the devastating impact of modern warfare tactics on civilian populations, particularly in populated areas. The report also emphasizes the disproportionate harm suffered by civilians, with the majority of those harmed being non-combatants, and the alarming trend of state actors being responsible for a significant portion of civilian casualties. These findings underscore the urgent need for concerted efforts to mitigate the impact of explosive violence and protect civilian lives in conflict-affected regions worldwide.

Table 1.2: Key Findings and Percentage of Violence Table

Key Findings	**Percentage**
Air-launched attacks	226%
Increase in civilian fatalities vs. 2022	122%
% of civilians harmed in towns/cities	90%
% of civilians harmed in populated areas	90%
% of civilian casualties from state actors	77%
% of violence incidents in populated areas	76%
Rise in explosive weapon use	69%
% of civilian fatalities from air strikes	67%
Ground-launched attacks	56%

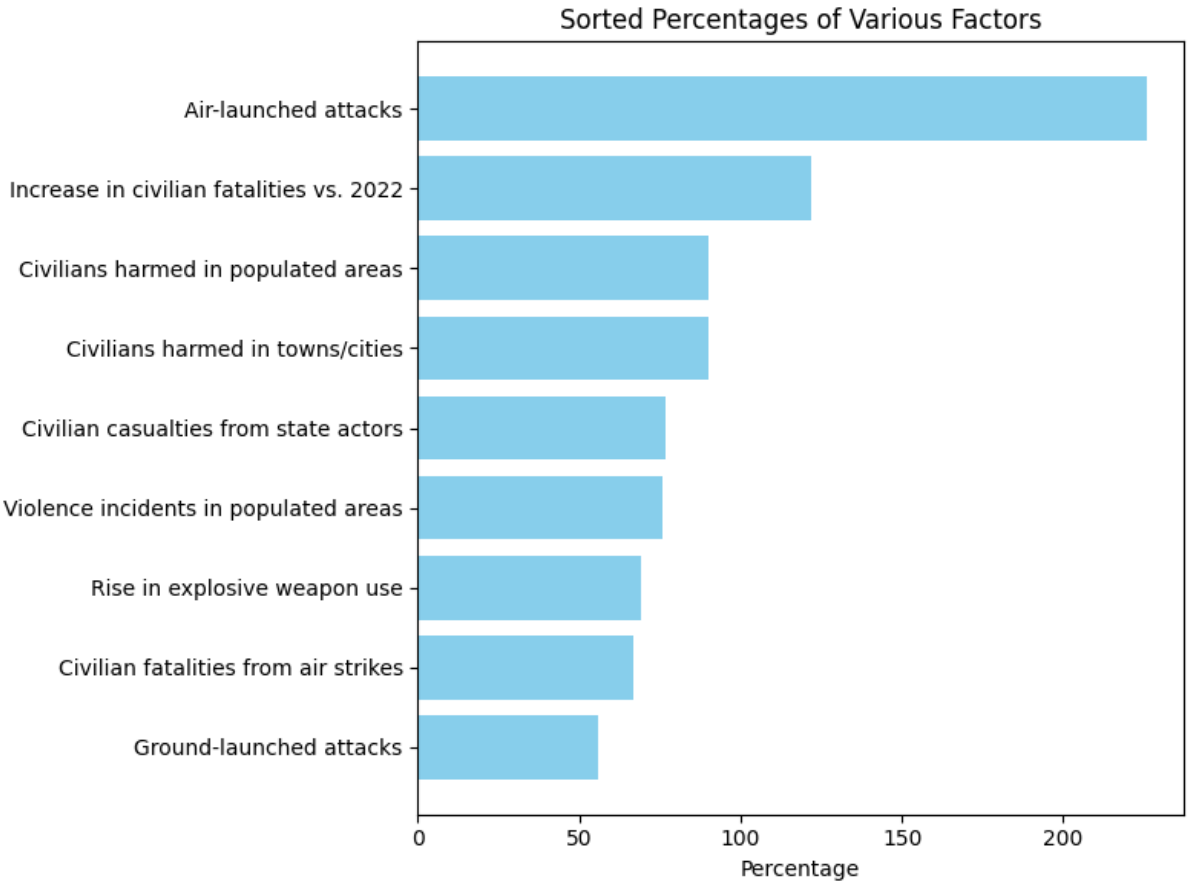


Figure 1.4: Graphical Representation of the Above Table

1.2 Objective and Scope

The development of robust and efficient models for violence detection from video is crucial for ensuring public safety and security. However, existing models often face challenges such as overfitting, complexity, and resource-intensive requirements, limiting their deployment in systems with limited specifications [5]. So the objectives of the project are:

- **Addressing Overfitting:** Overfitting is a common issue in existing systems and occurs when a machine learning model learns the training data too well but performs poorly on unseen data. So the aim is to develop a model that is less prone to overfitting, thereby improving its generalization capability [5].
- **Designing a Less Complex system:** The intention is to create a neural network that is less complex when compared to existing architectures. Beginning with a CNN-LSTM model, the aim is to refine it into an enhanced version that aligns with the objectives.
- **Enhancing Model Accessibility:** Due to the reasonably compact architecture of the proposed network, it can be trained and deployed on systems with reasonable specifications, thus increasing accessibility and cost-effectiveness.
- **Comparison with Highly Sophisticated Models:** Unlike highly sophisticated models that necessitate high-end systems for training and processing, the proposed network provides a more feasible alternative. Highly sophisticated models often demand extensive computational resources and may not be feasible for deployment on standard hardware. By prioritizing simplicity and efficiency, the proposed approach aims to reduce the gap between cutting-edge research and practical implementation.
- **Suitability for Edge Devices:** The model's reduced size and complexity make it well-suited for implementation on edge devices, including surveillance cameras, Internet of Things (IoT) devices, and more [6]. This enables efficient deployment in real-world scenarios, enhancing its practical utility.

Scope

The project aims to develop a robust system for violence detection and classification of corresponding actions as violent or not, utilizing machine learning and computer vision techniques to analyze videos. Its primary objective is to identify instances of physical confrontations or riots involving two or more parties. Upon detection, the system can be modified to notify relevant authorities promptly, facilitating timely intervention and resolution of disputes. By automating the process of detecting and categorizing violent incidents, the system seeks to contribute to the maintenance of law and order, ultimately enhancing safety and security in public spaces [7].

In line with enhancing safety and security, the project seeks to create a tool for monitoring and responding to potential threats. By deploying an intelligent surveillance system equipped with violence detection algorithms, the project aims to enable law enforcement agencies to stay ahead of security threats, thereby reducing the likelihood of violence and enhancing overall public safety.

The project's motivations are grounded in the prevention of criminal activities through timely intervention and the protection of vulnerable spaces. By intervening at the earliest stage, the project aims to disrupt criminal activities before they escalate, contributing to the reduction of crime rates such as assaults, vandalism, and riots. Moreover, by enhancing surveillance capabilities in high-risk areas, authorities can effectively monitor sensitive locations, reducing vulnerability to criminal activities and ensuring safety. Ultimately, the project seeks to promote community well-being by fostering trust, cohesion, and social harmony through the creation of safer and more secure environments. Through the empowerment of law enforcement agencies, the project aims to strengthen their capacity to protect and serve the community, enhancing public confidence to maintain safety and security.

1.3 Contributions

The project makes significant strides in the domain of public safety and violence prevention through a series of key initiatives. Primarily, the development of an advanced AI model tailored explicitly for detecting violence in video content [8]. This model incorporates techniques in deep learning and computer vision, ensuring precise analysis of video data to identify potential instances of violence. By leveraging AI algorithms, the proposed system offers an approach to threat identification and violence prevention, which is crucial for maintaining public safety.

Additionally, the project emphasizes the integration of data analysis capabilities, facilitating continuous monitoring of various data sources including surveillance cameras, social media platforms, and sensor networks [9]. This empowers law enforcement and security personnel to detect early signs of potential disturbances, enabling timely intervention to uphold public safety and order.

Together with technological advancements, ethical considerations, and accountability remain central to the project aim, prioritizing the development and deployment of the AI-powered violence detection system in accordance with ethical principles and individual rights [10]. The framework incorporates robust safeguards to mitigate risks such as privacy breaches and algorithmic biases, ensuring responsible and ethical operation.

Moreover, the project offers scalability and adaptability for flexible solutions adaptable to diverse contexts and environments. The modular design of the framework facilitates integration with existing public safety infrastructure and allows for customization to meet specific needs. This scalability and adaptability guarantee safer and more resilient communities.

The project endeavors to contribute to public safety and community welfare by leveraging AI technologies for detecting and preventing violence. It aims to offer authorities actionable insights and intervention strategies, envisioning safer public spaces conducive to peaceful coexistence [11]. The project seeks to cultivate safer communities and foster harmony in society, aspiring to create a more secure environment for future generations.

Chapter 2

Literature Review

2.1 Introduction

The literature survey explores the domain of violence detection and crowd management in surveillance videos, focusing on the application of deep learning techniques. These approaches are vital for ensuring public safety and security in various contexts, ranging from public spaces to high-security environments. By leveraging advanced algorithms and deep learning models, researchers aim to develop systems capable of automatically identifying violent incidents and analyzing crowd behavior in real time.

M. Shubber et. al.'s study on integrating machine learning and deep learning approaches [12] has yielded promising results in video violence identification. CNN was a reliable technique for extracting features from video frames, enabling accurate detection of violent behavior. Additionally, LSTM networks have been effective in capturing temporal dependencies in video sequences, overcoming issues like vanishing gradients, and leveraging time dimension information for improved analysis. However, a notable drawback of both CNNs and LSTMs is their reliance on supervised learning, where a large number of labeled training samples are required for training. Moreover, the computational demands of training these models can be substantial, requiring expensive hardware resources, which presents a practical challenge for widespread implementation in surveillance systems.

S. Lomlen's views on the exponential growth of AI [13] also present unprecedented opportunities for enhancing national security and safety across various domains. AI technologies offer the potential to revolutionize defense and military operations by enabling advanced threat detection, decision support systems, and autonomous capabilities. However, the integration of AI in high-stakes contexts like defense and military operations also brings forth significant challenges. One of the primary concerns is the lack of robustness, dependability, and safety of implementing AI methods in critical scenarios, where system failures or inaccuracies could have severe consequences. Additionally, ethical concerns arise regarding the decision-making capability of AI systems, especially in situations involving human lives and complex geopolitical dynamics. Addressing these challenges is crucial to harnessing the full potential of AI while ensuring the safety, security, and ethical integrity of its applications in national security contexts.

2.2 Existing Solutions

H.Gupta and Syed T.Ali's research [14] employs LSTM and Bidirectional LSTM (BiLSTM) models for violence detection. LSTM and BiLSTM are recurrent neural network architectures known for their ability to capture temporal dependencies in sequential data, making them well-suited for analyzing video sequences. These models are trained on annotated surveillance video data to learn patterns indicative of violent behavior. However, despite their effectiveness, these models come with computational challenges and may require significant resources for training and inference.

K.Aarthy and A.Alice Nithya's approach [15] focuses on violence detection using the Hockey dataset. To reduce computational costs, the study employs keyframe extraction, a technique that selects representative frames from video sequences. While this approach helps mitigate computational demands, it may struggle with generalization to new and diverse datasets. Additionally, suboptimal hyperparameter tuning could lead to decreased model performance, highlighting the importance of robust experimental design and optimization techniques.

M.Cheng et. al. have created and introduced the RWF-2000 dataset for violence detection [16]. This dataset consists of 2,000 video clips captured from real-world scenarios, providing researchers with a valuable resource for training and evaluating violence detection algorithms. Figure 2.1 shows the data collection pipeline of the RWF-2000 dataset. The Flow Gated Network, a unique architecture designed for this task, incorporates a self-learning pooling mechanism to enhance feature extraction from video data. However, the reliance on large amounts of labeled data for training poses a challenge, as annotating video content, especially violence-related material, can be labor-intensive and sensitive.

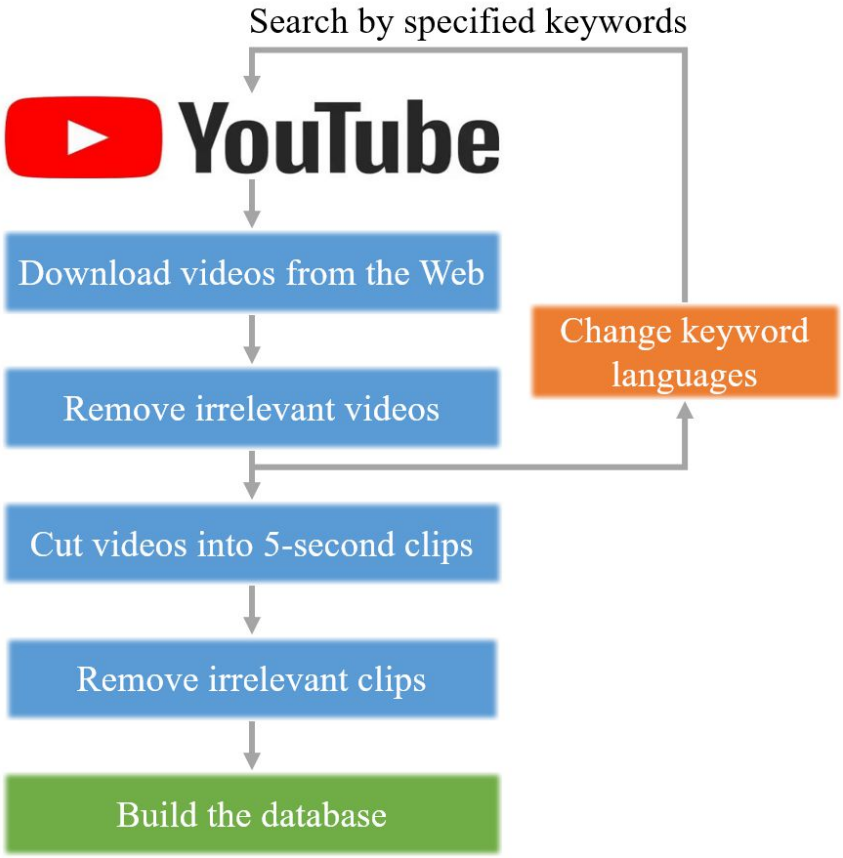


Figure 2.1: RWF-2000 Data Collection Pipeline

In the domain of crowd management, Y.Zuo et. al. have introduced the V3Trans-Crowd framework [17], which leverages a 3D visual transformer for analyzing crowd behavior in public spaces. This novel approach offers improved accuracy compared to existing methods, as demonstrated on the Crowd-11 dataset.

However, accurately classifying complex crowd behaviors remains a challenge, highlighting the need for further research into nuanced behavior recognition and classification techniques.

S.Vosta and K-C.Yow proposes a CNN-RNN structure for violence detection in surveillance camera feeds [18], combining Residual Network50 (ResNet) for feature extraction and ConvLSTM for anomaly detection. Unlike prior works on hand-crafted datasets, this study uses real-time surveillance feeds with diverse scenarios. It achieves promising results on the University of Central Florida (UCF) Crime dataset, surpassing models like Convolutional 3D (C3D) in Area under Curve (AUC). This research advances automated surveillance for enhanced security monitoring in public and private spaces.

A.Chauhan et. al. presents an overview of recent advancements in violence detection utilizing deep learning methodologies [19]. Studies such as Tiwari et al. [20], Bagga et al. [21], and Chauhan and Gupta [19] explore the application of CNN and hybrid models like the LHOGF algorithm combined with deep learning for real-time violence detection from Closed-Circuit Television (CCTV) footage. Despite achieving promising results, challenges such as processing delays in object detection remain, suggesting the need for further refinement and optimization of these models to enhance their real-time performance and accuracy.

R.G.Tiwari et. al. presents a novel approach for automated violence detection and classification in surveillance systems through a hybrid CNN-LSTM model [20]. By leveraging the strengths of CNN and LSTM networks, the proposed model achieves exceptional accuracy of 98.63%, surpassing both conventional machine learning methods and state-of-the-art deep learning systems. Through meticulous data collection and preprocessing techniques, the model was trained on a dataset comprising 11,043 images. The study underscores the effectiveness of the hybrid model in enhancing detection and classification skills for violent and nonviolent images in surveillance footage. Further research avenues include exploring additional hybrid architectures, optimizing hyperparameters, and expanding the model's capabilities to recognize a broader range of violent actions.

Y.Lyu and Y.Yang have presented a novel violence detection algorithm based on local spatio-temporal features and optical flow [22]. Unlike existing methods, this algorithm combines a physical contact detection algorithm with Harris 3D spatio-temporal interest point detection and optical flow to overcome computational challenges. By analyzing motion coefficients, it accurately identifies aggressive actions, distinguishing them from non-violent behaviors. Experimental results demonstrate the algorithm's effectiveness in real-time violence detection, particularly in scenarios involving multiple individuals.

Deepak K. et. al. presents a novel approach for violence detection using spatio-temporal autocorrelation of gradient-based features [23]. The study addresses the challenge of recognizing violent activities in crowded scenes, where traditional methods like trajectory analysis fail due to complex motion and occlusions. By focusing on simpler models, the proposed method, effectively extracts features and reduces computational complexity.

P.Sernani et. al. has introduced three deep learning-based models for automatic violence detection in videos and evaluate them on the AIRTLab dataset [24], specifically designed to challenge robustness against false positives. The schematic of all three models used on this dataset is given in Figure 2.2. The study emphasizes the importance of addressing misclassifications of friendly behaviors as violent actions. Transfer learning-based models exhibit stable accuracy across datasets, outperforming traditional approaches tested on benchmark datasets. Furthermore, 3D CNN-based models show superiority over 2D CNNs in processing spatio-temporal features, highlighting the potential of 3D architectures for violence detection. Despite challenges posed by imbalanced data, the AIRTLab dataset effectively tests the models' robustness. Future research aims to deepen comparisons between transfer learning and training from scratch approaches on various datasets. The paper underscores the significance of ongoing exploration in deep learning techniques for violence detection in surveillance videos.

Y. Su et. al. has introduced a new method for recognizing violent behavior by learning contextual relationships between related people from human skeleton points. The proposed Skeleton Points Interaction Learning (SPIL) module [25] aims to focus on the most relevant parts of skeleton points based on their features and position information.

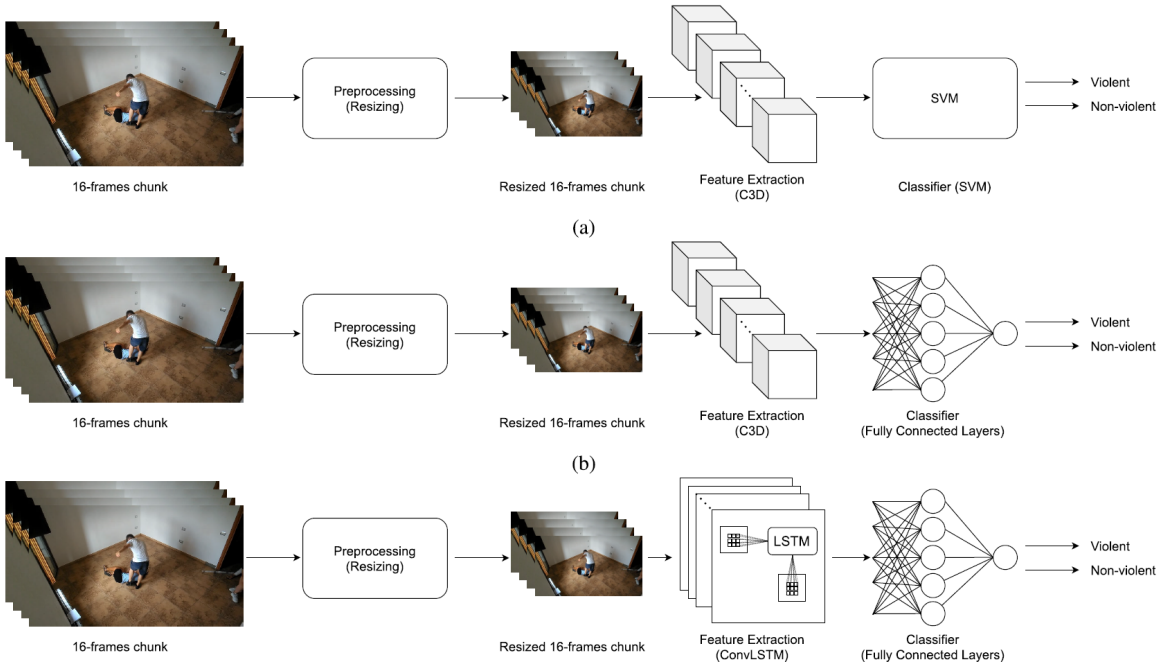


Figure 2.2: Schematic of the Three Models used on AIRTLab Dataset

Experimental results show that this model outperforms existing networks and achieves new state-of-the-art performance on video violence datasets. It achieves an accuracy of 89.3% on the RWF-2000 dataset and 95.5% on the Hockey dataset and Crowd Violence dataset.

Z. Islam et. al proposes a deep learning architecture that leverages innovative techniques to automatically detect violence from surveillance footage. The approach combines background suppressed frames and the difference of adjacent frames to highlight moving objects and capture motion, ultimately producing discriminative features for violence detection [26]. The proposed two-stream deep learning architecture leverages Separable Convolutional LSTM (SepConvLSTM) and pre-trained MobileNet for violence detection to effectively capture spatio-temporal features, distinguish between violent and non-violent actions and achieve state-of-the-art performance on benchmark datasets.
The schematic diagram of the proposed pipeline is shown in Figure 2.3.

G. Garcia-Cobo et. al. have introduced a novel deep learning architecture that accurately detects violent crimes by focusing on human skeletons and change detection in

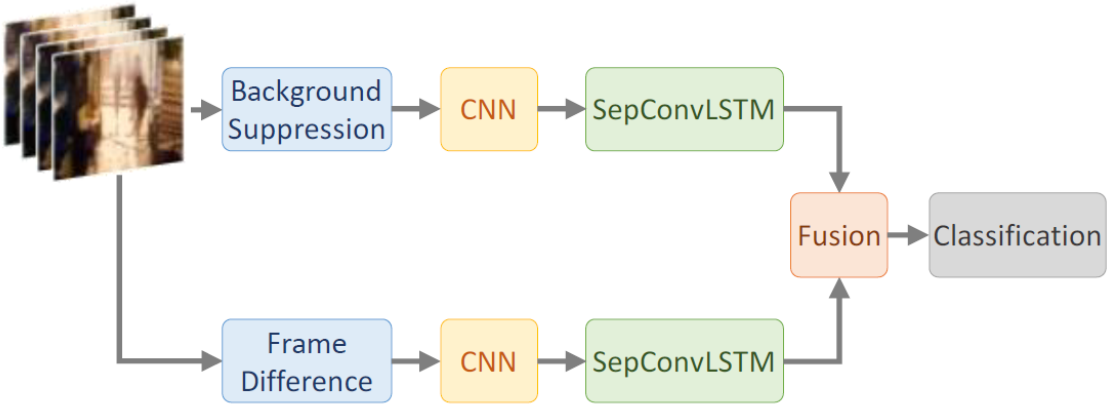


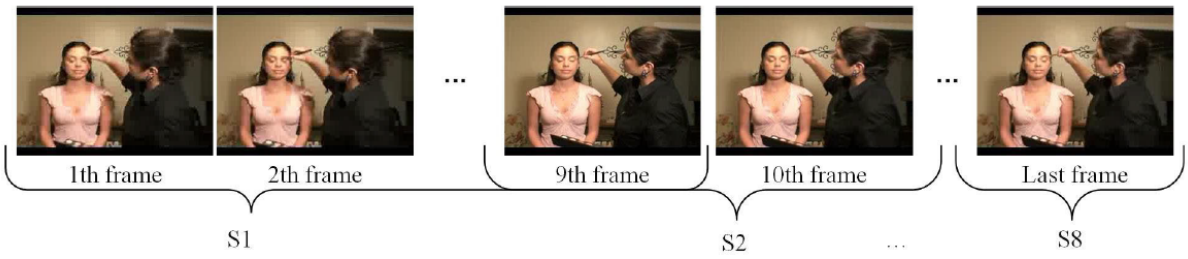
Figure 2.3: Schematic Diagram of SepConvLSTM Pipeline

surveillance footage [27]. By combining these key elements, the proposed method offers a promising solution for real-time crime detection. The model uses human pose extractors to capture spatial relationships in videos, incorporates change detection for identifying sudden movements, and combines these features using addition instead of multiplication. It also utilizes ConvLSTM for efficient processing of spatial and temporal information, enhancing violence detection accuracy.

H. Mohammadi et. al. discuss SSHA (Semi-Supervised Hard Attention) [28], a novel deep reinforcement learning model for video violence detection and localization. It addresses the need for scalable AI solutions in video surveillance by leveraging hard attention mechanisms to focus on relevant regions in video frames without relying on computationally expensive auxiliary features like optical flow or pose estimation. SSHA utilizes semi-supervised reinforcement learning, eliminating the need for localization annotations during training, and incorporates a pre-trained I3D backbone network to exploit temporal information. Experimental results demonstrate SSHA's state-of-the-art performance on multiple datasets, achieving accuracy rates of 90.4% on RWF and 98.7% on Hockey datasets with the RGB-only architecture, surpassing existing methods and showcasing the effectiveness of the hard attention approach.

R. Hachiuma et. al. proposes a novel deep neural network architecture called Structured Keypoint Pooling [29] to address limitations in conventional skeleton-based action recognition. With state-of-the-art accuracy and impressive speed on a single GPU, the proposed framework offers a solution to skeleton detection and tracking errors, poor variety of targeted actions, and person-wise and frame-wise action recognition challenges. The framework treats time-series key points as a 3D point cloud, allowing for sparse feature aggregation and mitigating errors associated with skeleton detection and tracking. This approach enables a broader range of actions to be targeted, including interactions with nonhuman objects, without relying on tracking algorithms. By leveraging this tracking-free architecture and sparse feature aggregation, the framework demonstrates improved performance in action recognition tasks compared to conventional methods, highlighting the effectiveness of the point cloud deep-learning paradigm.

Y. Qiao et. al. focuses on LaM-2SRN [30], a method designed to enhance local features and detect moving objects for action recognition. This paper explores a novel approach using a 3D CNN model to extract attention-enhanced spatiotemporal features for improved human action recognition. The LaM-2SRN model utilizes the traditional CAM (Class Activation Map) algorithm for visual attention in human action recognition by employing it as a target detection method to obtain optical flow information specifically from the human region. This helps in eliminating the influence of irrelevant optical flow information, such as background clutter, which can interfere with accurate action recognition. By focusing on the human body area using the CAM algorithm, the model can extract optical flow information relevant to the actions being performed, enhancing the discriminative features for better recognition accuracy. The input frames extraction network used in this research work and some examples of CAM-images, BE-images, and AE-images generated through the work is shown in Figure 2.4



(a) The Input Frames Extraction Network

(b) Examples of CAM-images, BE-images and AE-images

Figure 2.4: Activation and Saliency Maps

2.3 Summary

Various studies and research have led to the domain of violence detection in surveillance videos using machine learning and deep learning techniques. These approaches make use of advanced methodologies such as CNN, LSTM networks, and hybrid models to effectively extract spatio-temporal features from video frames. CNNs excel at feature extraction from visual data, enabling accurate identification of violent behavior patterns. Meanwhile, LSTM networks specialize in capturing temporal dependencies within sequential data, making them ideal for analyzing video sequences. Hybrid models combine the strengths of both CNNs and LSTMs, offering enhanced performance in violence detection tasks. Despite their success, these techniques encounter challenges, and addressing these challenges is crucial for refining existing models and exploring innovative approaches to further improve the accuracy and efficiency of violence detection in surveillance systems.

Deploying state-of-the-art violence detection models on edge devices like surveillance cameras faces several challenges as discussed before. Firstly, these models tend to be large, exceeding the storage and processing capacities of such devices. This limitation affects their feasibility for deployment in resource-constrained environments. Moreover, the extensive training time and high hardware demands further worsen this issue, making it challenging to train and run these models efficiently. Additionally, concerns about overfitting and the pursuit of enhanced classification accuracy contribute to increased computational requirements during both the training and inference phases. Consequently, implementing these models on edge devices becomes problematic due to the devices' limited computational resources and storage capacity, posing significant obstacles to accommodating the model parameters effectively.

Overall, the conclusion is that various machine learning and deep learning techniques can be employed for violence detection in surveillance videos, leveraging spatio-temporal features extracted from video frames. However, doing so in edge devices like surveillance cameras is challenging, thus limiting their feasibility and effectiveness in resource-constrained environments. To overcome these issues, there is a need to develop a shallow network with additional mechanisms for classification. For this, the team created a model less prone to overfitting while ensuring it fits within the memory constraints of edge devices like surveillance cameras. By optimizing the architecture to train efficiently on hardware with reasonable specifications, such a model can achieve effective violence detection while overcoming the challenges associated with memory limitations and hardware requirements.

Chapter 3

Proposed Methodology

3.1 Introduction

With this project, the team attempts to create a robust violence detection system for video streams using deep learning techniques. This holds significant potential for enhancing public safety and security in various environments, including surveillance, law enforcement, and crowd monitoring. The proposed model combines spatial and temporal feature extraction methods into the system aiming to improve accuracy and automate the detection process.

3.2 Data Collection

Effective violence detection systems rely on high-quality data for training and evaluation purposes. The process of collecting and curating data for violence detection entails several key considerations, including dataset size, diversity, and ethical considerations. In this section, an overview of the data collection methodology employed in violence detection research is explained.

Data collection begins with the identification of relevant sources, including surveillance videos, movies, television shows, and online platforms. Various sources were identified and researched to obtain the most suitable dataset for the best performance of the Model. The team embarked on a mission to enrich the dataset by staging meticulously choreographed violent scenes [31], which were then recorded to inject diversity into the dataset. Additionally, the team meticulously combed through CCTV footage [32] from various locations, including the college premises and neighboring shops and hotels, in search of real-life instances of violence with the intention of a violence detection framework that solely relies on visual features, eliminating the need for audio input [33] [16]. However, despite the efforts, the dataset obtained fell short in size, revealing the need for alternative approaches to expand it.

Efforts were made to stage violent scenes like collection of videos in [34] involving team members but encountered challenges due to the constraints posed by costumes, particularly the college uniform, which hindered the accurate identification of individuals and detection of their actions. This made the team shift its strategy towards exploring data available online. Thus, the team navigated through various avenues, adapting their approach to overcome obstacles and ensure the dataset's adequacy for training the model. Public datasets were the most favorable option, as they offered a wide range of data samples from diverse sources, ensuring a comprehensive representation of real-world scenarios for training the model.

In the publicly available datasets, several datasets can meet the violence detection process standards, but only a few can perform well. This is due to the scarcity of such datasets with proper labels for a systematic supervised learning algorithm. Here are some of the popular datasets:

1. **UCF-Crime Dataset:** The UCF-Crime dataset contains various types of violent and non-violent criminal activities captured from surveillance videos. It includes over 13,000 video clips annotated with violence labels, making it a valuable resource for violence detection and action recognition research. The content of this dataset is shown in Figure 3.1.

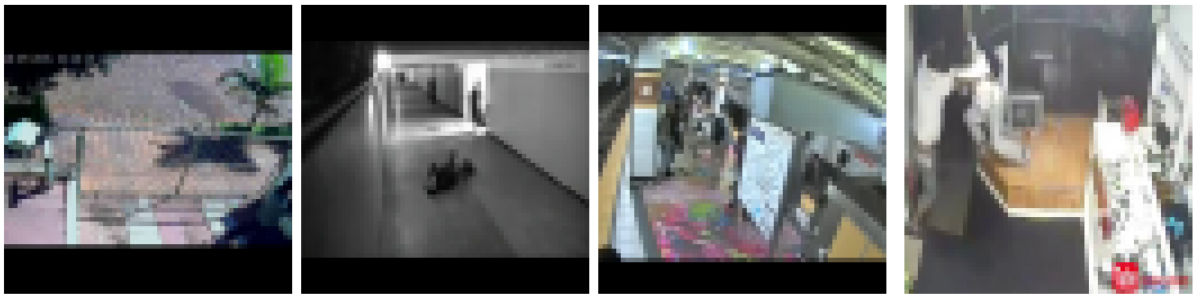


Figure 3.1: Contents of UCF-Crime Dataset

Some challenges faced by this dataset are:

- The dataset primarily consists of surveillance videos capturing criminal activities, which may not adequately represent the diversity of public violence scenarios.
- This dataset is designed for general anomaly detection whereas anomalies in action may or may not include violent activities.
- The dataset may suffer from imbalanced class distributions, with a higher prevalence of non-violent activities compared to violent ones.
- Videos in this dataset have a duration of up to 10 minutes, Hence, detecting violent activities from a long video is very difficult [16]

2. **Movie Dataset:** The Movie Dataset is a collection of video clips extracted from movies, providing researchers with a diverse range of visual content for various applications, including violence detection. While this dataset offers rich and varied scenes depicting a wide array of actions and behaviors, including violent and non-violent interactions, it also presents several challenges for violence detection research:

   - Movies span a wide range of visual content with varying levels of violence. Models must be robust enough to detect violent actions across different contexts and styles.

- The dataset may contain scenes with varying levels of resolution, lighting conditions, artificial elements, and camera viewpoints, impacting the visual representation of violent actions in the real world
- The Movies dataset may possess copyright issues with data being extracted from movies and may require the expenditure of money for alleviating copyrights.

3. **Hockey Fight Dataset:** The Hockey Fight Dataset is a curated collection of video clips extracted from hockey games, specifically focusing on instances of fights between players [15]. These clips are annotated to indicate the presence or absence of fights for research in violence detection and action recognition. The overview of the Hockey Fight dataset is given in Figure 3.2.

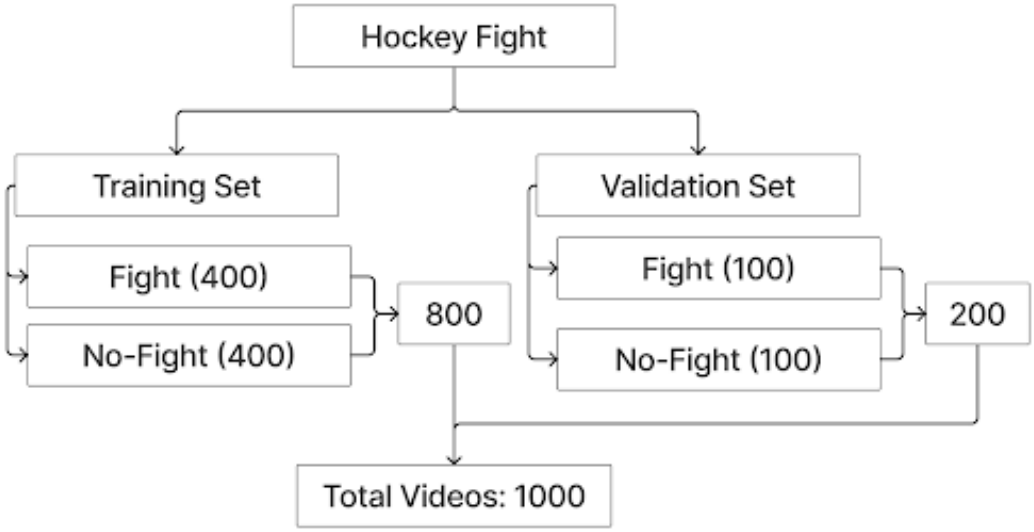


Figure 3.2: Hockey Fight Dataset Overview

Here are some of the advantages of the Hockey Fight Dataset.

- The white color in the background of the hockey dataset videos likely provides a clear contrast with the players and other objects in the scene. This facilitates background separation, making it easier to isolate and track the players' movements and actions accurately.

- The compact nature of the hockey dataset, consisting of 1000 videos, offers several advantages for research and analysis. Despite its relatively small size compared to larger datasets, it remains valuable due to its focused scope and curated content.
- High-quality video recordings with clear resolution and minimal motion blur ensure that the players' movements are captured accurately, enhance the interpretability of the dataset, visualization of player actions, such as skating, passing, shooting, and defending,

Nonetheless, this dataset also has some drawbacks including a lack of diversity of videos [16], Variability of duration, intensity, and multiple number of participants, which make it harder for algorithms to extract useful information.

4. **RWF-2000 Dataset:** The RWF-2000 dataset is a widely used benchmark dataset for action recognition and violence detection research. It contains video clips extracted from movies, television shows, and YouTube videos, depicting various types of violent and non-violent behaviors. The dataset consists of 2,000 video clips, with 1600 clips used in the training set, 800 videos each for class Fight and Non-Fight, and the remaining 400 used in the validation set, 200 videos each for class Fight and Non-Fight [16]. This dataset, whose overview and contents are displayed in Figure 3.4 and Figure 3.3 respectively, showed superior performance when it comes to the detection of violence properly and accurately.

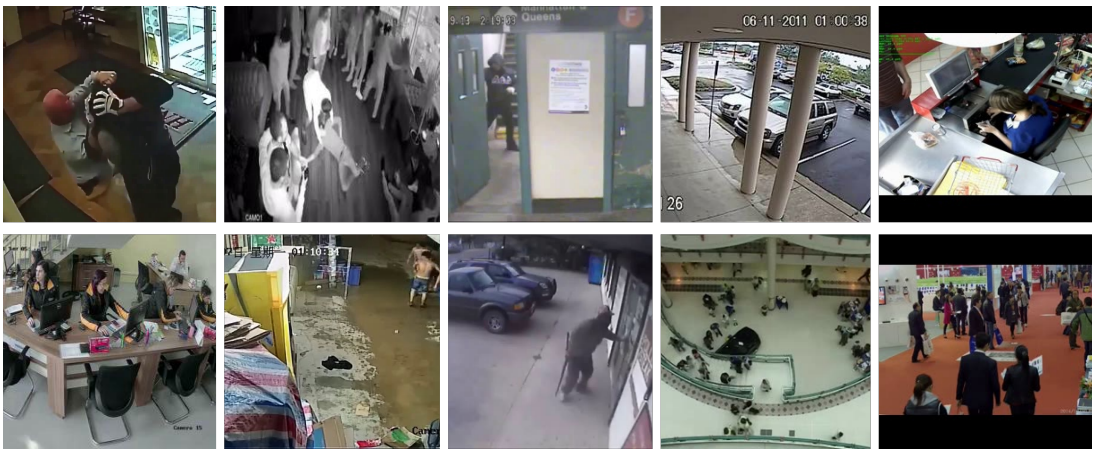


Figure 3.3: Contents of RWF-2000 Dataset

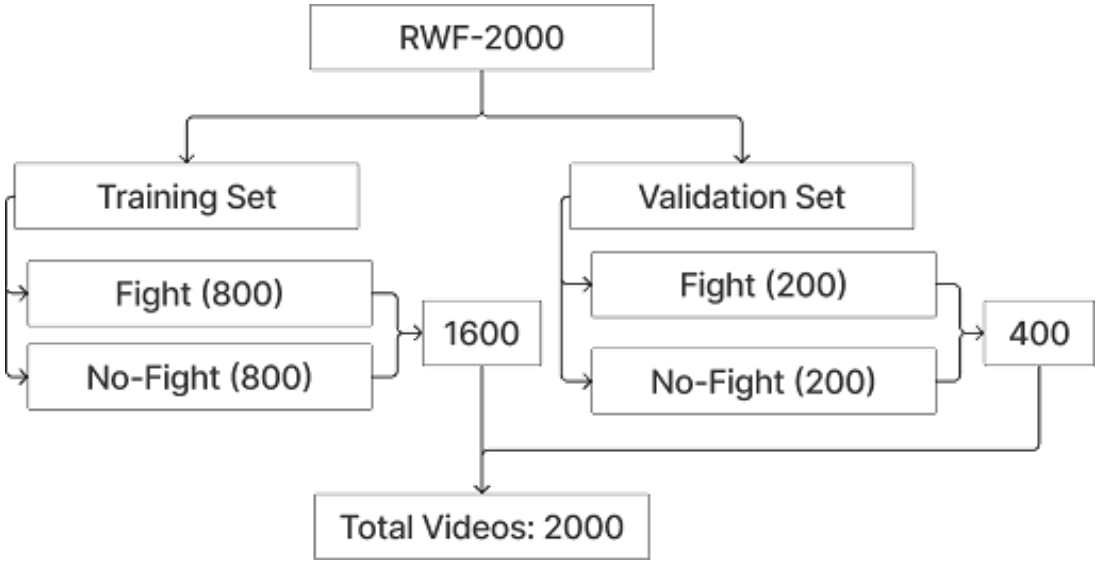


Figure 3.4: RWF-2000 Dataset Overview

Some of the advantages of RWF-2000 over other datasets for violence detection are in Table 3.1 and Table 3.2.

Table 3.1: Advantages of RWF-2000 Dataset

No.	**Feature**
1	The RWF-2000 dataset contains over 2,000 video clips, providing a substantial volume of data for training and evaluation.
2	Each video clip consists of 150 frames (30fps x 5 sec), ensuring equal duration and size across the dataset.
3	The dataset is balanced with 1000 clips each for Fight and Non-Fight classes, avoiding class imbalances.
4	Arranged with an 80% train (1600 samples) and 20% test (400 samples) split, ideal for optimal training and testing.
5	Widely recognized as a benchmark in violence detection research, it serves as a standard reference for algorithm evaluation.

The resolution distribution of the RWF-2000 dataset is given in Figure 3.5.

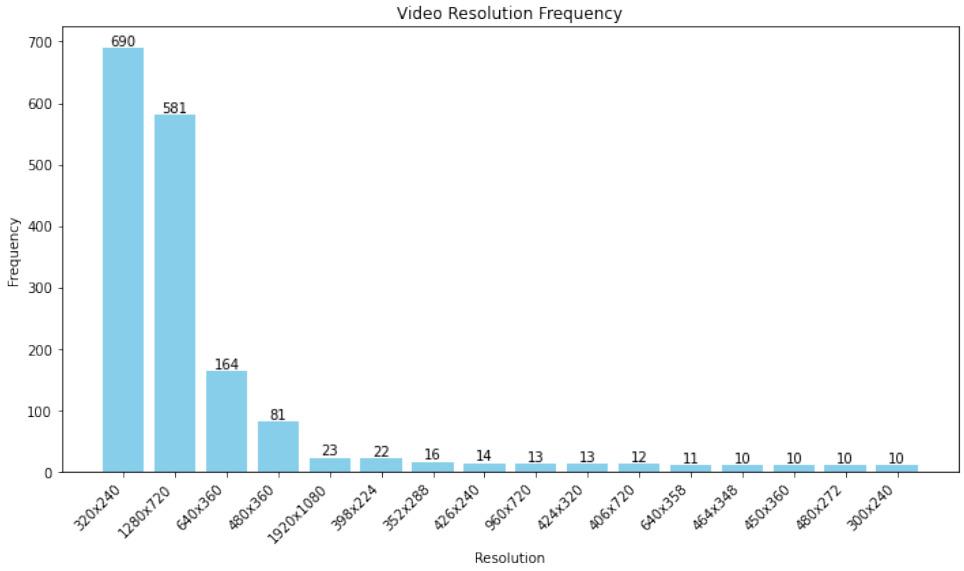


Figure 3.5: RWF-2000 Video Resolution Distribution

Table 3.2: Disadvantages of RWF-2000 Dataset

No.	Disadvantage
1	Quality from YouTube varies widely, introducing inconsistencies that can affect model performance.
2	Use of specific keywords may miss some types of violence, leading to a biased dataset.
3	Videos of real people used without consent raise ethical concerns and privacy issues.
4	Models may overfit the types of violence in the dataset and won't generalize well.
5	Lack of contextual information can make it difficult to distinguish between real and staged violence.
6	Potential legal issues related to copyright when using YouTube videos for research or commercial purposes.
7	Inaccuracies in video labeling can mislead training and affect detection system performance.

3.3 Dataset Preprocessing

The Preprocessing begins with the transformation of raw video data, which undergoes initial preprocessing steps to enhance interpretability and facilitate downstream analysis. The data preprocessing encompasses the following steps:

- Renaming the videos based on their content, such as labeling fight-related videos as "Fight_1..." and "No_Fight_1...," for videos with no fight, which ensures better organization and understanding of the dataset.
- Subsequently, the videos are subjected to frame extraction that decomposes the continuous video stream into individual frames. Each video's frames are then saved in a corresponding folder, named after the original video, simplifying data management and retrieval as shown in Figure 3.6.
- Following frame extraction, the frames undergo resizing to a standardized format of 224x224 pixels with three color channels (224x224x3). This resizing ensures compatibility with most pre-trained deep learning models and facilitates efficient processing. While resizing maintains the essence of the video content, preserving three color channels balances the trade-off between training time, accuracy, and computational speed.
- Transforming data into the .npy format simplifies data management, data retrieval, and streamlining operations with efficient memory mapping. This optimization enhances read and write speeds, optimizing resource utilization for smoother training and inference processes.

In the context of data handling, .npy files offer advantages over individual image files by consolidating multiple video frames into a single numpy array file. This simplifies the process of accessing and managing data, as users only need to interact with the .npy file instead of iterating through numerous image files placed inside different folders. The process of storing and searching in NumPy files is shown in Figure 3.7.

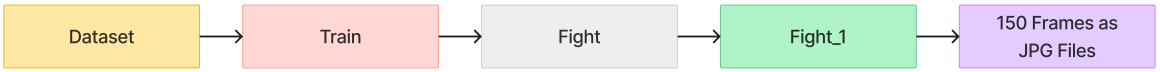


Figure 3.6: Storing and Searching in Normal Files

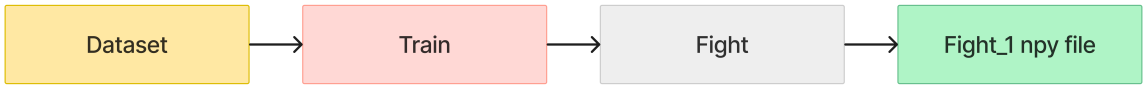


Figure 3.7: Storing and Searching in NumPy Files

Moreover, .npy files are stored in a binary format, resulting in smaller file sizes compared to collections of image files, which further enhances efficiency during data manipulation and storage. Figure 3.8 shows how the memory is mapped as a path to these .npy files.

```
'C:/Users/ajayt/OneDrive/Desktop/Main P/full_data\\val\\fight\\fight_100.npy': array([1., 0.]
'C:/Users/ajayt/OneDrive/Desktop/Main P/full_data\\val\\fight\\fight_101.npy': array([1., 0.]
'C:/Users/ajayt/OneDrive/Desktop/Main P/full_data\\val\\fight\\fight_102.npy': array([1., 0.]
'C:/Users/ajayt/OneDrive/Desktop/Main P/full_data\\val\\fight\\fight_103.npy': array([1., 0.]
'C:/Users/ajayt/OneDrive/Desktop/Main P/full_data\\val\\fight\\fight_104.npy': array([1., 0.]
'C:/Users/ajayt/OneDrive/Desktop/Main P/full_data\\val\\fight\\fight_105.npy': array([1., 0.]
```

Figure 3.8: Memory Mapping Printed as Path to the NumPy Files

The combination of these preprocessing steps sets the stage for subsequent feature extraction and model training, enabling the development of a robust violence detection system. By transforming raw video data into a structured and standardized format, the project lays the foundation for effective analysis and interpretation. Additionally, the systematic organization of data enhances the scalability and reproducibility of the violence detection pipeline.

3.4 Model Building

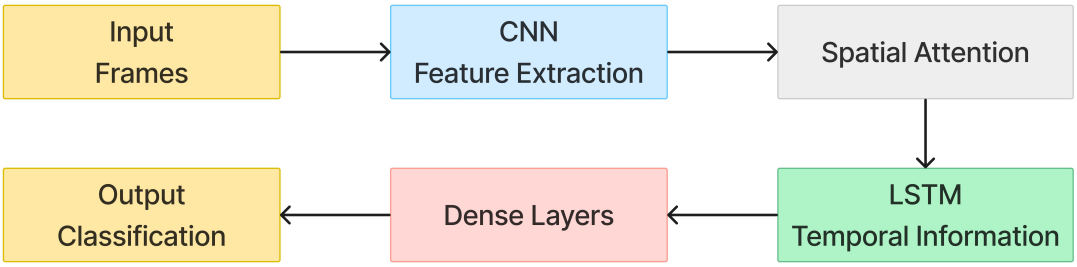


Figure 3.9: High-Level Overview of the Proposed System

The design of the model started from a simple idea of *how humans look and understand a violent situation*. Consider an example of a violent activity happening some distance away from a person, the person tends to focus on speech, disturbance, and movements. After some milliseconds the focus will be shifted to more detailed things in the view, like if a weapon is used, the emotional state of people surrounding the situation (anger, fear, and disgust) within a short duration of time the person will understand the situation and act accordingly. The team concentrated on building this type of response using deep learning methods, The initial focus of the person can be implemented using a special type of deep learning architecture called CNN which is an abstraction of human perception. To make the CNN focus on minute and detailed items present in the view the team used a simple attention mechanism called *Spatial Attention*, which identifies the brightest(regions with high RGB values) and most averaged parts(repeated across many video frames, like quick movements) and makes a map out of it and compares with the original frames and returns the attended region which makes the important features stand out more during the network's learning process. But this Spatial Attention only helps to understand what is in a frame and where to focus. To act like a human, the network needs to know how the information changes over time, for this another deep learning architecture called LSTM was used. LSTMs are a type of Recurrent Neural Networks (RNN), that are designed to handle data in sequences. It has memory cells and other gates, which can maintain its state over time, effectively allowing the network to remember past information for an extended period. This is key to understanding temporal patterns and dependencies(time-based patterns).

Together CNN, Spatial Attention, and LSTM simulate the working of a human brain when the person is near a violent situation. But this can change with the situation (in technical terms, the data), as this biological response pattern is a general pattern for action recognition in humans, as for the proposed network architecture.

The high-level overview the proposed system consists of 6 different phases, each of which has a different deep learning architecture used. The diagrammatic representation of these phases can be seen in Figure 3.9. Details of each of these steps from data preprocessing to final output are described as follows.

3.5 Detailed Description of the System

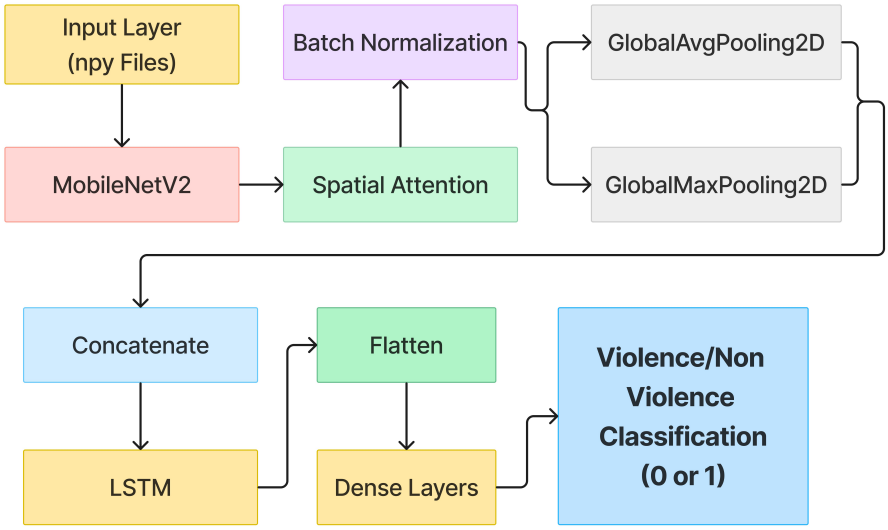


Figure 3.10: Block Diagram of the Proposed System

Figure 3.10 shows the architecture of the proposed system which is a simple but effective model construction that consists of different components working together to classify whether a video contains violence or not. The approach is a single-stream approach where the data sequentially goes through each layer of the model.

3.5.1 Frame Extraction and Sampling:

The preprocessing pipeline for video input begins with frame extraction, breaking down continuous streams into individual frames for subsequent analysis. Converting videos into sequential frames enables processing at the frame level, essential for analysis by computer vision models, facilitating action recognition and event detection in video data.

Resizing these frames to a standardized 224x224x3 format enhances computational efficiency and accelerates learning processes while reducing complexity. Standardizing resolutions ensures consistency across the dataset, enabling models to learn invariant features and improving generalization performance on unseen data.

Converting frames into NumPy's .npy format optimizes storage and access, leveraging the format's efficient binary representation for faster read and write operations than traditional image file formats. Converting to .npy format streamlines data handling, reducing complexity and enabling efficient memory mapping for faster read and write operations, optimizing resource utilization during training and testing.

To maintain temporal context while managing computational resources effectively, a uniform sampling strategy is employed. This approach involves selecting frames at regular intervals, ensuring a balance between computational efficiency and the retention of essential information.

3.5.2 Feature Extractor CNN

The rich representations inside each frame of a video will be captured by feature extraction CNN models, in this case, the team has used a lightweight and fast pre-trained model named MobileNetV2, which is designed for the implementation on edge devices and provides a quick output. It is suggested in [35] that effectiveness and accuracy can be improved by the usage of pre-trained models.

The MobileNets uses a special kind of CNN called Depthwise Separable Convolution, it has a minor change from traditional CNN, as it splits the computation into two steps: depthwise convolution and pointwise convolution, the first applies a single convolution filter to each of the input channel and the later is used to create a linearly combined output of the depthwise convolution, these actions reduce the total number of multiplications required in computation than a traditional CNN. In the proposed model the 20 layers on the bottom of MobileNet were frozen, as there is no need to train the model to capture high-level details like edges, bright spots, etc.

But going to the top all other layers are trainable because it is required to capture minute details present in video frames. All of these operations are done using a special kind of layers called TimeDistributed layers, which do the corresponding operations to all the frames passed to the network in a batch. The code implementation is given below.

```
from tensorflow.keras.applications import MobileNetV2
base_model = MobileNetV2(include_top=False, weights='imagenet',
    input_shape=input_shape[1:])
base_model.trainable = True
for layer in base_model.layers[:-20]: # Freeze last 20 layers
    layer.trainable = False
frames_features = TimeDistributed(base_model)(inputs)
```

3.5.3 Spatial Attention Mechanism :

The output of MobileNetV2 is a group of feature maps that contains the spatial and temporal dimensions. The Spatial Features [36] extracted from video frames capture static information, while temporal features [37] represent dynamic changes over time. The model needs special attention towards the spatial information, to focus on important parts of the frames, they are passed through a Spatial Attention module. This helps the model to improve its ability to capture important visual features associated with violence. The spatial attention works in a simple manner, where the maximum and average values across the frame batch are calculated and concatenated along the channel axis.

This is then passed through a single filter CNN to make an activation map(shown in results and discussions) having values between 0 and 1. Finally, a multiplication operation is done against the feature map with the original input tensor element-wise, this multiplication operation basically acts as a gating mechanism where values close to 1 in the attention map allow the corresponding features in the input tensor to pass, and values close to 0 will be blocked. The code used for spatial attention is given below.

```
class SpatialAttention(Layer):
    def __init__(self, **kwargs):
        super(SpatialAttention, self).__init__(**kwargs)

    def build(self, input_shape):
        self.conv2d = Conv2D(1, (7, 7), activation='sigmoid',
           padding='same')
        super(SpatialAttention, self).build(input_shape)

    def call(self, inputs):
        max_pool = tf.reduce_max(inputs, axis=-1, keepdims=True)
        avg_pool = tf.reduce_mean(inputs, axis=-1, keepdims=True)
        concat = Concatenate(axis=-1)([max_pool, avg_pool])
        attention = self.conv2d(concat)
        return Multiply()([inputs, attention])
```

3.5.4 Batch Normalization:

Following the spatial attention mechanism and learning in general the distribution of activations in each layer may shift, leading to slower convergence and degraded performance. Batch normalization is a useful technique that can be applied to stabilize the activations, which helps improve the convergence and training speed of the model.

3.5.5 Feature Pooling & Concatenation:

The output from batch normalization is then split into two branches. One branch is dedicated to GlobalAveragePooling2D, and the other to GlobalMaxPooling2D. These pooling operations reduce the size of feature maps and output combined information about the features. The results from both pooling operations are concatenated to capture comprehensive spatial information from the video frames.

3.5.6 LSTM Network:

The concatenated feature vector resulting from the pooling operations is then inputted into an LSTM network. LSTM networks are a type of RNN specifically designed for modeling sequential data and are proficient at capturing temporal dependencies and long-term patterns within sequences. In the context of video analysis, LSTM enables the model to effectively encode temporal information and identify complex patterns that extend over multiple frames. This is made possible by the presence of memory cells and gating mechanisms inside the LSTM.

3.5.7 Dense Layer & Classification:

Following the LSTM layer, the output(2D Tensor) is flattened using a Flatten layer(to 1D vector) and is passed through a collection of dense layer with dropout layers fitted between them to prevent overfitting. The final dense layer is followed by a softmax activation function to classify the output into two classes: violence or non-violence(labelled as 'Fight' and 'No Fight').

The key layers or components of the proposed system with its important functions are listed below in Table 3.3. The summary of the model architecture is shown in Table 3.4 while the model definition generated using the Keras Plotting utility is shown in Figure 3.11.

Table 3.3: Key Layers/Components of Proposed System with Functions and Importance

Layer/Component	**Function**	**Importance**
Frame Extraction	Converts video into sequential frames.	Essential for analyzing actions frame by frame.
Frame Resizing	Standardizes frame size to 224x224x3.	Ensures uniform input dimensions for consistent processing.
NumPy Conversion (.npy)	Optimizes frame storage and access.	Enhances data handling speeds and training efficiency.
Uniform Sampling	Selects frames at regular intervals.	Balances computational load and info retention.
Feature Extractor CNN (MobileNetV2)	Extracts features using depthwise separable convolutions.	Extract spatial features within frames.
Spatial Attention Mechanism	Highlights important spatial features in frames.	Focuses model on significant areas, improving detection accuracy.
Batch Normalization	Stabilizes activations, improve training efficiency.	Ensures faster, more stable training convergence.
Feature Pooling & Concatenation	Reduces dimensionality and combines features.	Captures comprehensive spatial information from frames.
LSTM Network	Captures temporal dependencies and patterns across frames.	Essential for recognizing sequences of actions that indicate violence.
Dense Layer & Classification	Processes features to classify the video as 'violence' or 'non-violence'.	Makes the final decision using learned features to detect violent behavior.

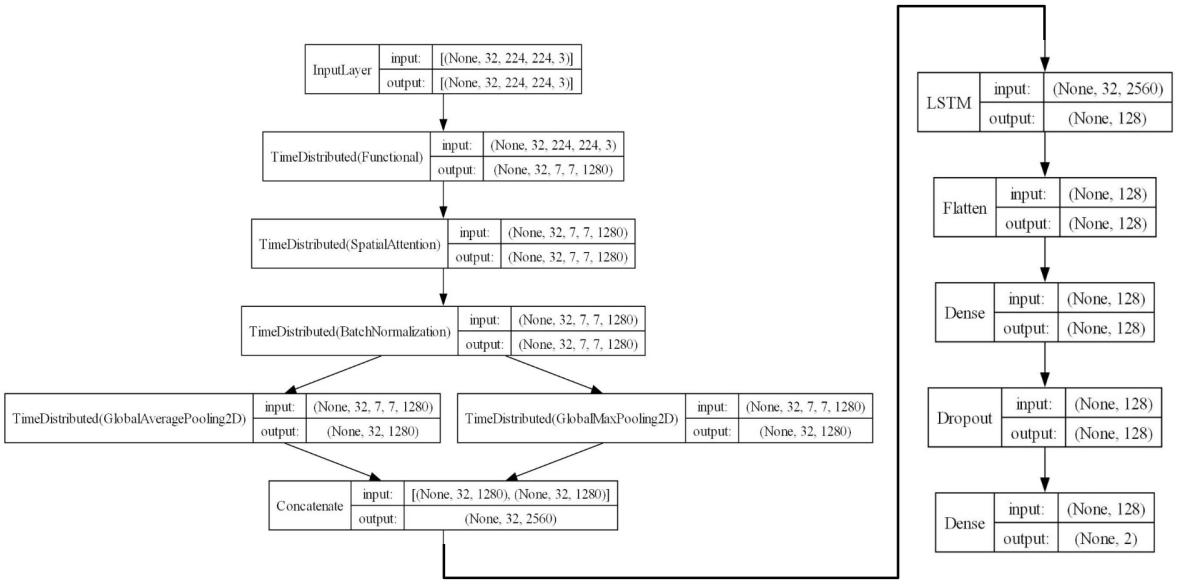


Figure 3.11: Keras Plotting Utility's Output of Model Definition in Code

Table 3.4: Summary of Model Architecture

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 20, 224, 224, 3)	0	[]
MobileNetV2*	(None, 20, 7, 7, 1280)	2,257,984	['input_1[0][0]']
Spatial Attention*	(None, 20, 7, 7, 1280)	99	MobileNetV2*
Batch Normalization*	(None, 20, 7, 7, 1280)	5,120	Spatial Attention*
GlobalAveragePooling2D*	(None, 20, 1280)	0	Batch Normalization*
GlobalMaxPooling2D*	(None, 20, 1280)	0	Batch Normalization*
Concatenate	(None, 20, 2560)	0	GlobalAvg/MaxPooling2D*
lstm (LSTM)	(None, 128)	1,376,768	['concatenate[0][0]']
flatten (Flatten)	(None, 128)	0	['lstm[0][0]']
dense (Dense)	(None, 128)	16,512	['flatten[0][0]']
dropout (Dropout)	(None, 128)	0	['dense[0][0]']
dense_1 (Dense)	(None, 25)	3,225	['dropout[0][0]']
dropout_1 (Dropout)	(None, 25)	0	['dense_1[0][0]']
dense_2 (Dense)	(None, 10)	260	['dropout_1[0][0]']
dropout_2 (Dropout)	(None, 10)	0	['dense_2[0][0]']
dense_3 (Dense)	(None, 2)	22	['dropout_2[0][0]']

*TimeDistributed Layer | Total Parameters: 3,659,990

3.6 Coding Practices

To make the model as per the objectives and ideas the team used top-of-the-line technologies available to make the proposed deep learning networks. And orchestrated the whole process of designing and testing the model using the industry-grade productivity software named Notion. Other tools used to make the model are outlined in the Table 3.5

Table 3.5: Programming and Deep Learning Framework Details

Item	**Description**
Programming Language	Python [38], a programming language focused on readability and easiness of implementing ideas.
Python version	3.10.11, the version that supports most libraries
Deep Learning Framework	TensorFlow 2.10 [39], Keras 2.10 [40]
Other Libraries	OS, Numpy, OpenCV, Pandas, Matplotlib, and Seaborn
DL Architectures used	CNN, LSTM, GlobalAveragePooling2D, GlobalMaxPooling2D, Concatenation Layer, Flatten Layer, Dense Layer, and Dropout Layer.

Table 3.6: Model Training Parameters

Parameter	**Value**
Mixed precision	ON, float32
Batch Size	4
Num_Frames	20
Height x Width of each frame	224 x 224
Num_Channels	3 (RGB)
Num_Classes	2 (Fight, No Fight)
Learning Rate	0.001 to 0.0009
Loss Function	categorical_crossentropy
Optimizer	Adam
Output Activation	Softmax

The parameters were selected after a concise run with different arrangements of these parameters. Some other parameters were found to be working in certain types of problems in literature, like a thumb rule, when using categorical cross entropy as the loss function, put the num_classes as 2 and output activation as Softmax, or else use num_classes as 1, output activation as Sigmoid and the loss function should be binary cross entropy. The paper introduced by V Deepa Et. al [41] suggest using Adam Optimizer with a learning rate of 0.01 for this type of classification problem to get an effective learning curve without much overfitting.

Since the project involves processing videos there is a requirement for high memory in Read Only Memory(ROM), Random Access Memory(RAM), and Graphics Processing Unit-Video Random Access Memory (GPU VRAM). During the training phase, the data has to be stored in RAM and when the training starts the data has to be moved to GPU memory. During the training, there will be another kind of memory usage like making intermediate tensors during backpropagation, and Frameworks like TensorFlow have their own memory overheads for managing computations. But If the system memory or graphics memory is low, then the program will hit an Out-Of-Memory(OOM) Error, thereby halting the execution of the program.

To curb this problem, there are three methods the team has implemented: Frame Sampling, Adding Data Generators, and Setting Incremental Memory Growth for GPU.

1. Frame Sampling - Using a sampled set(uniform sampling, random sampling, cluster sampling) of frames rather than using all the frames of a video as the input to the model.

2. Data Generators - Similar to generators in Python which yields the specified amount of data during execution. Data Generators are also used to make the required training data and input data augmentation on the fly, especially in tasks like video processing.

3. Incremental GPU Memory Growth - Frameworks like TensorFlow do not have built-in automated memory usage limits, they tend to allocate arrays and tensors to the

total available memory upfront to the execution. But there are options to specify the usage of memory incrementally, one of them is setting memory growth for GPUs in TensorFlow. This can be implemented in code by,

```
gpus = tf.config.experimental.list_physical_devices('GPU')
if gpus:
    try:
        for gpu in gpus:
            tf.config.experimental.set_memory_growth(gpu, True)
    except RuntimeError as e:
        print(e)  # Memory growth must be set before GPUs have
           been initialized
```

• List of Callbacks: Perform specific actions at various stages of training. Some of the callbacks used in the code are shown below.

```
checkpoint_cb = ModelCheckpoint(checkpoint_path, save_best_only=
   True, verbose=1)
reduce_lr_cb = ReduceLROnPlateau(monitor='val_loss', factor=0.2,
   patience=3, min_lr=0.001, verbose=1)
early_stopping_cb = EarlyStopping(monitor='val_loss', patience
   =20, verbose=1, restore_best_weights=True)
csv_logger = CSVLogger(os.path.join(res_path, "training_log.csv")
   , append=True)
```

epoch	accuracy	loss	lr	val_accuracy	val_loss
0	0.542500019	0.703077137	0.001	0.855000019	0.565554082
1	0.725000024	0.570079744	0.001	0.824999988	0.515035689
2	0.772499979	0.532961667	0.001	0.824999988	0.48389855
3	0.816250026	0.484389126	0.001	0.839999974	0.483640075
4	0.821250021	0.467398763	0.001	0.824999988	0.486711055
5	0.814999998	0.452991396	0.001	0.860000014	0.482051313
6	0.852500021	0.394777387	0.001	0.850000024	0.408922881
7	0.857500017	0.404281437	0.001	0.894999981	0.337867498
8	0.877499998	0.36955905	0.001	0.845000029	0.404371619
9	0.863749981	0.432847589	0.001	0.904999971	0.302209646
10	0.86500001	0.337587386	0.000904837	0.894999981	0.331467777

Figure 3.12: Snapshot of Training Log

- ModelCheckpoint: helps to keep checkpoints at a specific training point which helps to recover that point if subsequent training gets halted by some unforeseen circumstances. This is useful when training a large model or when there is a need to add more information to an existing model.

- ReduceLROnPlateau: Changes the learning rate when the specified metric (here, validation loss) is not improving. Learning Rate changes at the end of the previous epoch

- EarlyStopping: Used to stop training when the specified metric (here, validation loss) goes out of range (means, the training curve and validation curve go away from each other)

- CSVLogger: Stores a piece of training information like epoch count, training accuracy, testing accuracy, training loss, and validation loss onto a CSV file. Figure 3.12 is a snapshot of CSVLogger which shows how the results are arranged in the Comma Separated File (CSV) file.

• Learning Rate Scheduler: Keeps specified initial learning rate for the first five epochs (in this case) and then it will make changes to the learning rate. Learning Rate is changed at the beginning of the current epoch.

```
def scheduler(epoch, lr):
    if epoch < 5:
        return lr
    else:
        return lr * tf.math.exp(-0.1)
lr_scheduler = LearningRateScheduler(scheduler)
```

Chapter 4

Results and Discussions

4.1 Model's Prediction on Unseen Data

The Figure 4.1 and Figure 4.2 show the predictive power of the model when given by a .npy file as input, during this testing phase the layers like Batch Normalization and Dropout Layers will be off. The output(prediction) of the model will be an average of calculations of the input frames in the model, as the model goes on to attend all the frames to make a representation of the input file.

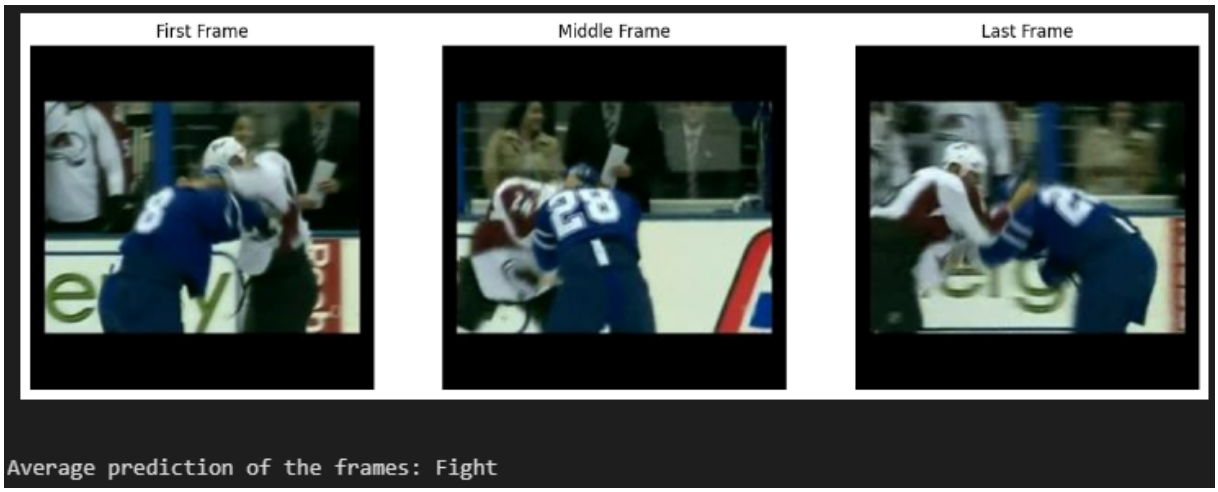


Figure 4.1: Model Prediction on a 'Fight' Video

Figure 4.2: Model Prediction on a 'No-Fight' Video

4.2 Performance Evaluation

In this section, the performance of the proposed model is checked using various metrics to assess its effectiveness and reliability in achieving the intended outcomes. The primary objective of the project was to propose a lightweight system, that does not overfit on the data and trains in reasonable hardware resources. Though with the training most of these objectives were covered, there is a need to analyze the model in-depth to understand whether the model is robust, fast, and effective to unseen data.

- Dataset Used: Hockey Fight Dataset
- Evaluation Metrics: Accuracy, Precision, Recall, and F1 Score.
- Testing methodology: The model will be tested on a separate part of the dataset called the testing set, which consists of 200 Videos under the labels Fight and No_Fight.

4.2.1 Accuracy-Loss Graph

The Accuracy-Loss Graph is an essential visual tool that illustrates the model's performance across its training and validation phases. Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases the model has gone through. Loss, on the other hand, quantifies the error between the predicted values and the actual values, which shows the model's precision during training. Therefore, these graphs help in understanding the model's learning curve and pinpointing any overfitting or underfitting issues.

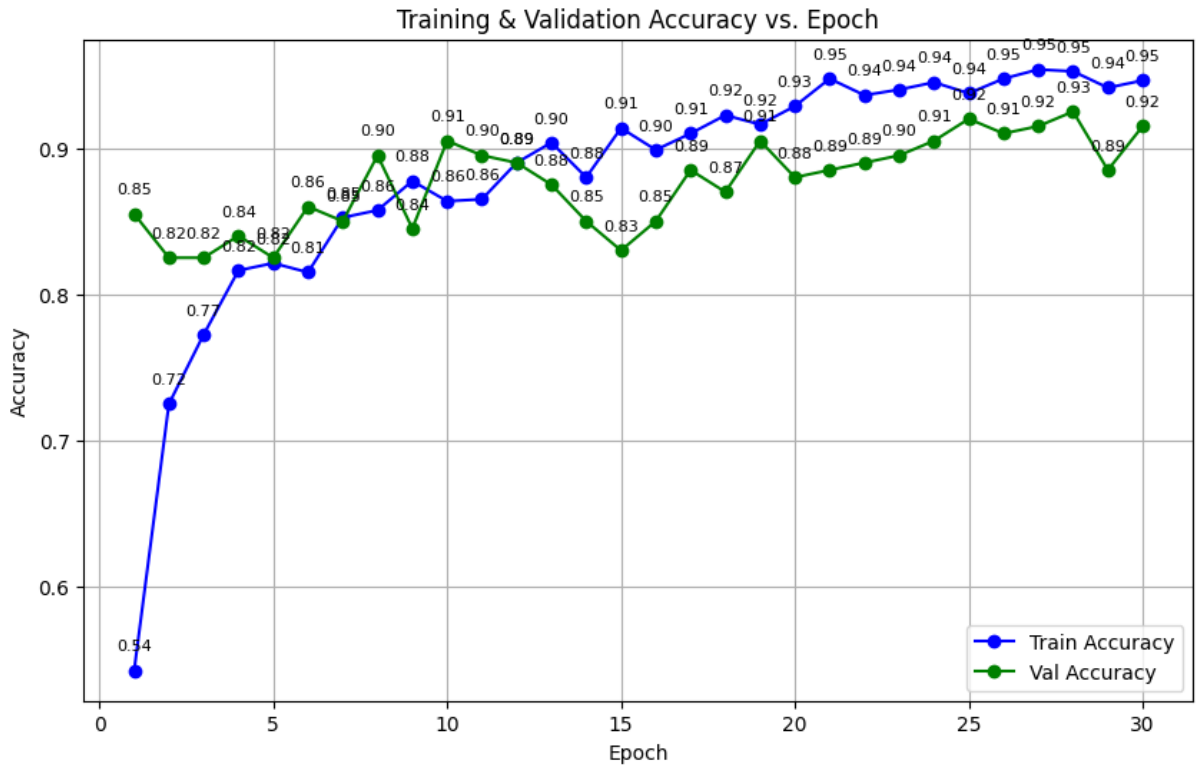


Figure 4.3: Accuracy versus Epoch Graph

From these Figures 4.3, 4.4, 4.5 it is clear that the model has shown good performance in training, as well as testing. The space between the accuracy curves shows whether the model is overfitting or not, but on keen observation, it is clear that there are no serious signs of overfitting. Also, the troughs and crests are due to the presence of noise and resolution problems of the videos in the dataset.

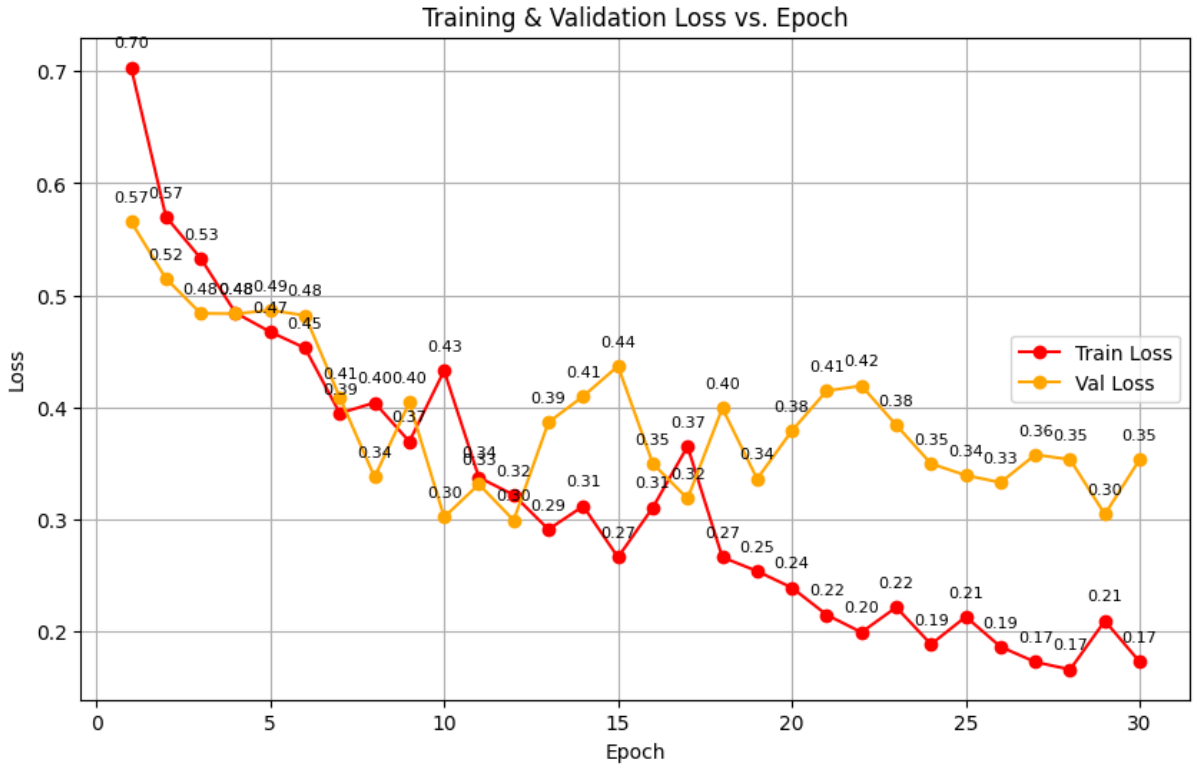


Figure 4.4: Loss versus Epoch Graph

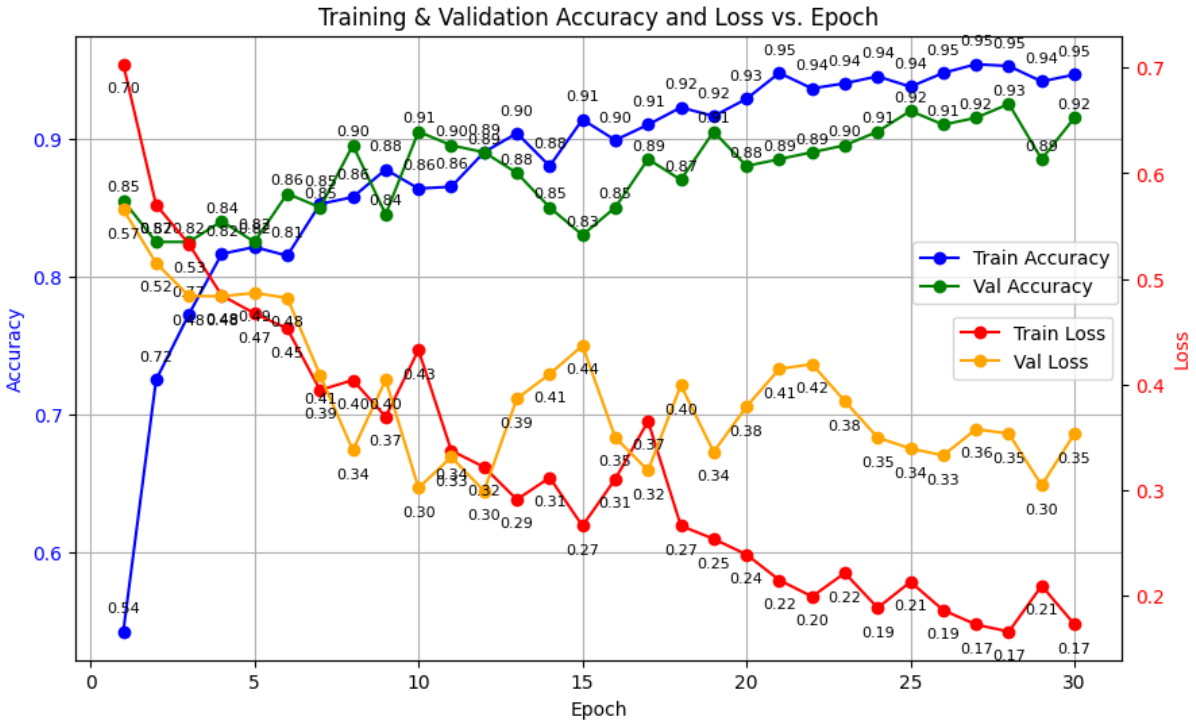


Figure 4.5: Accuracy and Loss versus Epoch in a Single Graph

4.2.2 Confusion Matrix

The Confusion Matrix is a tabular representation of the classification power of the deep learning models. It encodes the values to find other metrics like Precision, Recall, and F1 Score. The confusion matrix of the proposed model is given in Figure 4.6

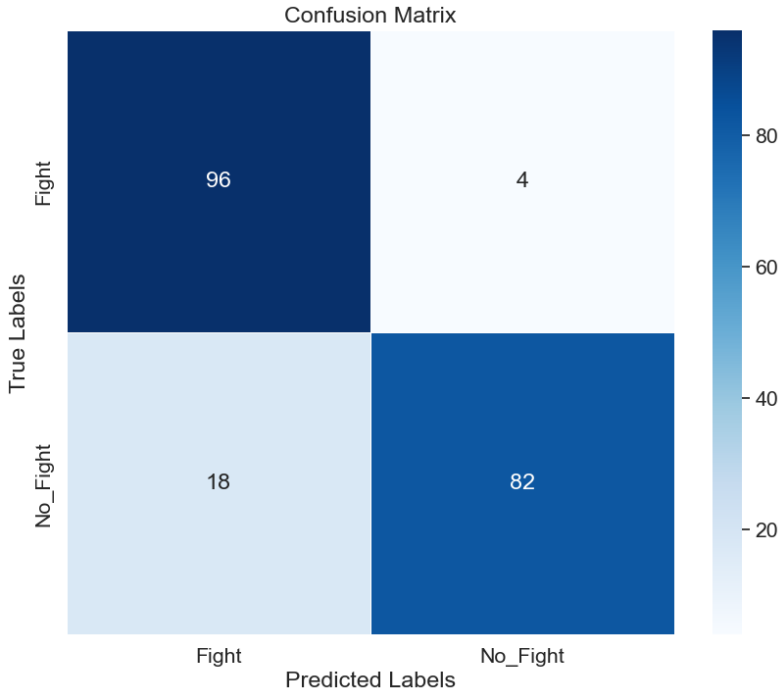


Figure 4.6: Confusion Matrix of the Proposed Model

Confusion Matrix Analysis:

1. True Positive (TP): 96 (Top-left cell) - Correct predictions where the actual class is "Fight" and model predicted, "Fight".

2. False Positive (FP): 18 (Bottom-left cell) - Times model incorrectly predicted "Fight" when the actual class was "No_Fight".

3. False Negative (FN): 4 (Top-right cell) - Times model incorrectly predicted "No_Fight" when the actual class was "Fight".

4. True Negative (TN): 82 (Bottom-right cell) - Correct predictions where the actual class is "No_Fight" and the model also predicted "No_Fight".

4.2.3 Accuracy

Accuracy is a common performance metric in deep learning, particularly useful for classification problems. It measures the proportion of true results (both true positives and true negatives) to the total observations(true positives, false positives, true negatives, and false negatives). Accuracy is computed as in the general equation 4.2.1,

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2.1}$$

$$\text{Accuracy (Training/Testing)} = \frac{96 + 82}{96 + 82 + 4 + 18} = \frac{178}{200} = 89\%$$

4.2.4 Precision

In violence detection systems, precision is crucial because it measures how many of the detected instances of violence are actually correct. High precision means that when the system identifies an event as violent, there is a high probability that it truly is violent. With high precision, unnecessary panic or resource wastage is minimized. Precision is computed as in the general equation 4.2.2,

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2.2}$$

$$\text{Precision (Fight)} = \frac{96}{96 + 4} = \frac{96}{100} = 96\%$$

$$\text{Precision (No\_Fight)} = \frac{82}{82 + 18} = \frac{82}{100} = 82\%$$

4.2.5 Recall

Recall is equally critical because it measures the system's ability to detect all actual violent events captured in the video data. In the context of public safety and security, a high recall rate ensures that no violent incidents go undetected. This is particularly vital in surveillance systems used in public areas like schools, malls, or public transportation, where failing to detect an act of violence can have severe consequences. Recall is computed as in the general equation 4.2.3,

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2.3}$$

$$\text{Recall (Fight)} = \frac{96}{96 + 18} = \frac{96}{114} \approx 84.2\%$$

$$\text{Recall (No\_Fight)} = \frac{82}{82 + 4} = \frac{82}{86} \approx 95.3\%$$

4.2.6 F1 Score

The F1 score becomes particularly important when balancing the trade-offs between precision and recall. In violence detection, it is often crucial to maintain a balance where both false positives (non-violent acts labeled as violent) and false negatives (violent acts not detected) are minimized. This balance is essential because both types of errors can have serious implications—false positives can lead to unnecessary interventions, while false negatives might allow violent situations to escalate. The F1 score provides a single metric that helps optimize the model during training and tuning to ensure an effective balance between identifying true violence and reporting non-violent activities. Precision is computed as in the general equation 4.2.4,

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.2.4}$$

$$F1(\text{Fight}) = 2 \cdot \frac{0.96 \cdot 0.842}{0.96 + 0.842} \approx 89.7\%$$

$$F1 \text{ (No\_Fight)} = 2 \cdot \frac{0.82 \cdot 0.953}{0.82 + 0.953} \approx 0.882 \text{ or } 88.2\%$$

Table 4.1: Training and Testing Metrics

Criteria	**Metric**	**Value (%)**	
		Training	**Testing**
Graph	Accuracy	95.00	92.00
Testing	Accuracy	89.00	89.00

Table 4.2: Fight vs No Fight Metrics

Metric	**Value (%)**	
	Fight	**No Fight**
Precision	96.00	82.00
Recall	84.20	95.30
F1-Score	89.70	88.20

In addition, the overall result values that were computed are written in Table 4.1 and Table 4.2. The accuracy metrics (given in Table 4.1) are different for graph and model testing, this is because of the usage of automatic mechanisms to save the best keras model file based on validation accuracy which is explained in the coding practices section.

4.3 Spatial Attention Maps

Spatial Attention Maps helps to highlight regions of interest within images or data, aiding in the interpretation and understanding of complex visual information. The original image acts as the input, and the spatial attention map overlays above the original image to produce an output as shown in Figure 4.7.

(a) Original Image (b) Attention Map (c) Resultant Attention Map

Figure 4.7: Steps in Generating Attention Map

(a) Original Image (b) Attention Map

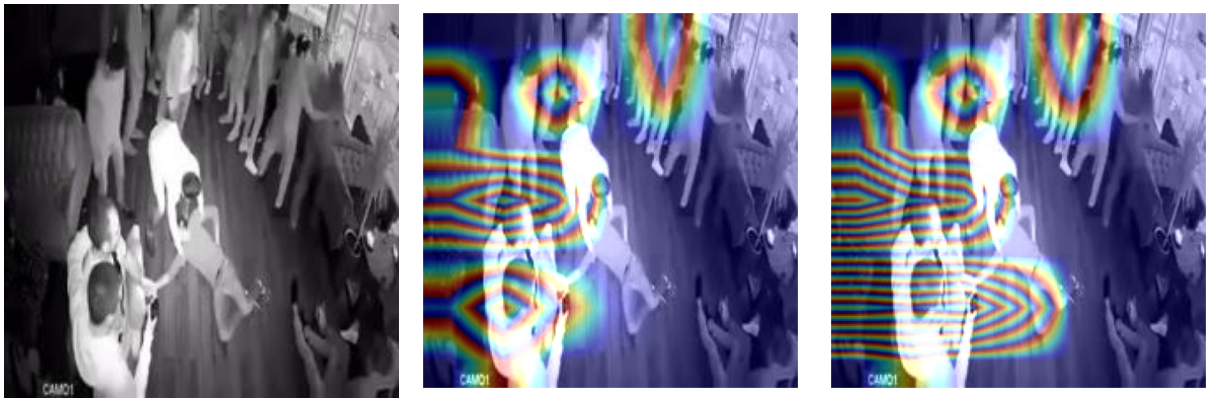
Figure 4.8: Correct Attention Map

(a) Original Image (b) Attention Map

Figure 4.9: Wrong Attention Map 1

(a) Original Image (b) Attention Map

Figure 4.10: Wrong Attention Map 2



(a) Original Image (b) Learning Attention(A) (c) Learning Attention(B)

Figure 4.11: Learning Happening in Attention

- Purpose: Spatial Attention Maps are used to visually highlight areas within an image that a computational model focuses on when making decisions or predictions. This can help researchers and engineers understand what parts of an image are marked important by the model.

- Process: The generated attention map is overlaid on the original images. These maps are color-coded or heat-mapped layers as shown in Figure 4.7(b) that indicate the 'attention' a model gives to different parts of an image when analyzing it. The correct attention map outputs are shown in Figure 4.8 and the wrong ones in the 4.9 and Figure 4.10. In addition, Figure 4.11 shows how the learning takes place in attention maps over each iteration.

- Educational and Diagnostic Use: These attention maps cannot be used for improving the algorithms by tweaking them, but they serve an educational purpose. They can help explain to students and professionals how certain algorithms interpret visual data and what features the network prioritizes. This can be classified as a part of ExplainableAI (XAI)

4.4 Discussions

Training a deep learning model without any manual feature engineering to perform well on complex data like real-world fight and violence data is inherently a difficult procedure. In order to make it possible, one way is to make a very deep network with complex architecture that can absorb all the intricacies and nuances in the training data, another way is to do manual feature engineering like annotating the dataset to tell the model where the violence or fight is occurring. Both of these require heavy manual work to make it possible.

Some of the reasons that fuel the difficulties are that the data were mostly captured from CCTV visuals which has many problems like low resolution, distant and different camera angles views. To make a well-performing and lightweight model(as per the objective) we have to input a lot of data, do validation, and incorporate many precautions to handle edge cases like overfitting, OOM errors, and other problems that are explained in the proposed methodology chapter. But still, the model has managed to overcome these difficulties which were justified by the results of the testing.

Effectively, the performance metrics like the accuracy-loss graphs and confusion matrix, show insight into the effectiveness and efficiency of our model. Precision shows a measure of the model's exactness by indicating the proportion of positive identifications that were actually correct. Recall provides insight into the model's completeness, reflecting its ability to identify all relevant instances in the dataset. Finally, the F1 Score combines precision and recall into a single metric that captures both the false positives and false negatives. It is particularly useful when the balance between precision and recall is important.

Chapter 5

Conclusion

In the domain of human activity detection, particularly in identifying instances of violence, combating overfitting presents a significant challenge, even with the utilization of large and deep layered models [5]. This challenge primarily occurred due to the complex nature of the model itself and the extraction of irrelevant features during training. To address this issue, the project undertook the development of a network with a less complicated architecture compared to existing models.

By employing a simpler architecture, the proposed network can be effectively trained on systems with standard specifications, eliminating the necessity for high-end computing resources. Moreover, we integrated additional mechanisms such as spatial attention, global average pooling, and global max pooling to augment the model's capability to extract relevant features while reducing the risk of overfitting.

The accuracy of the proposed model aligns closely with existing systems, the noteworthy accomplishment lies in the significant reduction of overfitting. Furthermore, the model architecture was carefully designed to prioritize computational efficiency and resource utilization [42]. This strategic approach enables seamless deployment on edge devices with limited processing power, such as surveillance cameras, facilitating precise classification between violent and non-violent activities.

Finally, the proposed approach has made substantial progress towards achieving the project's objectives of developing a lightweight, faster, and compact model capable of accurately distinguishing between violent and non-violent activities. Through a combination of simplified architecture, feature enhancement mechanisms, and a focus on computational efficiency, the proposed model represents a promising solution for real-world deployment in scenarios where resource constraints are a primary concern.

5.1 Future Scope

The future of violence detection using AI holds significant promise and potential for advancements in several key areas:

- **Audio-Visual Cues Integration:** By combining visual and auditory information, these systems can better understand violent incidents, improving accuracy and reliability. The model that detects child violence from voice variation is an example. [43]
- **Predictive Classification:** Implementing a method that can identify the likelihood of violence on a per-frame basis before the violence has occurred and caused any harm by localizing the regions of the frame or the violent people in view. [26]
- **Security and Law Enforcement Alert:** Violence detection systems can be integrated with emergency services, such as police dispatch centers or emergency call centers. When a violent incident is detected, the system can automatically generate an alert and relay relevant information to dispatchers.
- **Privacy-Preserving Solutions:** As privacy concerns continue to grow, there is a need for AI-driven violence detection solutions that respect individuals' privacy rights. Future research may explore privacy-preserving techniques, such as federated learning [44] and differential privacy, to ensure sensitive data remains secure and confidential.

- **Detecting Violent Activities Involving Weapons:** Continuous evaluation of additional standard datasets should encompass a broader range of violent activities, including those involving weapons [14], which present unique challenges for detection algorithms due to their concealment and potential variability in appearance.

- **Spectral Imaging:** Spectral imaging [45] offers a unique set of capabilities for violence detection such as usage of the IR spectrum, allowing for the identification of weapons, bloodstains, chemical residues during very low lighting conditions, complete darkness, or even through obstacles.

5.2 Limitations

While the project has made significant progress in addressing overfitting and improving computational efficiency for violence detection, there are several potential limitations to consider:

- **Generalization:** Despite efforts to reduce overfitting, there may still be instances where the model fails to generalize well to unseen data or variations in real-world scenarios. This could lead to misclassifications or reduced accuracy in certain situations.

- **Resource Constraints:** While the model aims to be computationally efficient, it may still require a certain level of processing power and memory, which could be prohibitive for deployment on extremely resource-constrained edge devices.

- **Scope Limitation:** While your model excels in detecting physical violence, it may not be equipped to identify other forms of violence, such as weaponized violence

- **Environmental Factors:** Variations in lighting conditions, camera angles, and environmental clutter (e.g., crowded spaces, occlusions) may impact the model's performance, leading to decreased accuracy or increased false positives/negatives in violence detection.

Chapter 6

Project Activities and Outreach

6.1 ICDICI-2024 Conference Submission

The 5th International Conference on Data Intelligence and Cognitive Informatics (ICDCI 2024) invites global researchers, scholars, and industry professionals to explore the intersection of data intelligence and cognitive informatics. With a focus on understanding information processing systems, the conference aims to foster the development of advanced cognitive informatics technologies. The proof and reply email of the conference is shown in Figures 6.1 and 6.2 respectively.

- **Conference Name:** 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI 2024)
- **Publication:** Institute of Electrical and Electronics Engineers (IEEE)
- **Conference Dates:** 18^{th}-20^{th} November, 2024
- **Date of Submission:** 2^{nd} May, 2024

- **Acceptance Intimation:** 12th September, 2024
- **Conference Website**: *https://www.icdici.com/*
- Notably this Conference was listed by the official conference searching site of the IEEE.
- The selected paper immediately after the conference presentation will be submitted to IEEEXplore Digital Library.

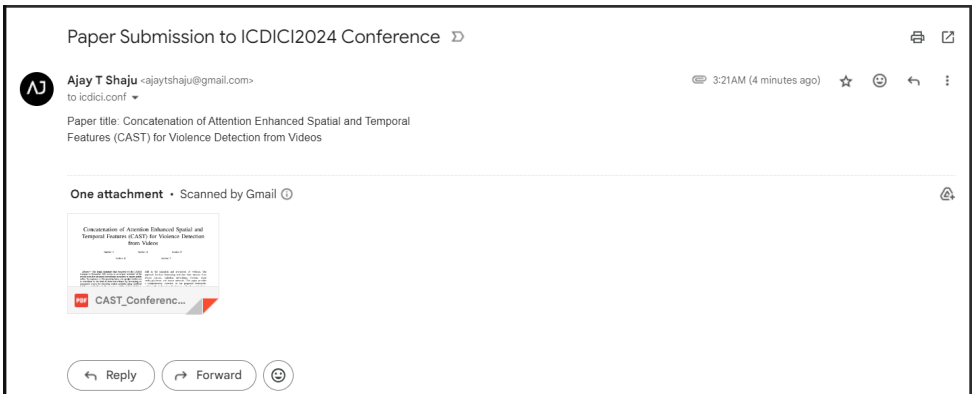


Figure 6.1: Proof of Research Paper Submission to ICDICI2024 Conference

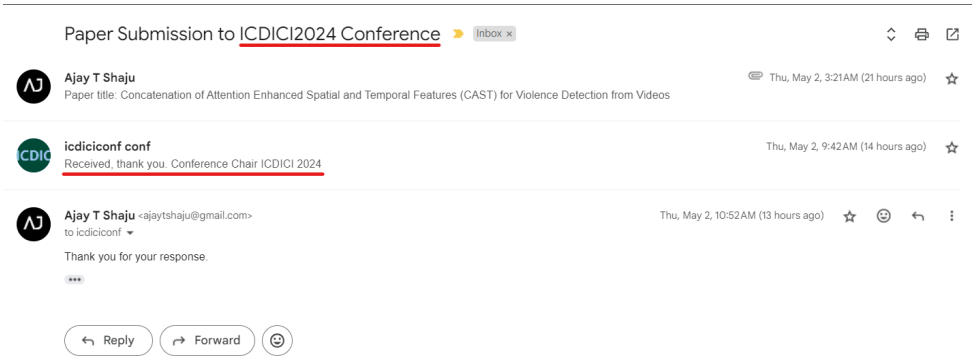


Figure 6.2: Reply - 'Paper Received' from ICDICI2024 Conference

6.2 Carmel College Project Competition

Figure 6.3: Project Team Members Attending Project Competition at Carmel College of Engineering, Alappuzha held on 29th April 2024. From left: Ajay T Shaju, Emil Saj Abraham, Vishnuprasad KG, and Justin Thomas Jo

The team members participated in a project competition hosted at Carmel College of Engineering, Alappuzha Figure 6.3. The project presentation emphasized key features including the lightweight model, its scope, and notably the stampede incident at CUSAT served as the project's motivation, which seems to be easily understood by the judges of what the team has done. This experience provided an opportunity for all team members to showcase and exchange ideas with like-minded individuals, receiving valuable insights in return. Moreover, it enabled the team members to appreciate and learn from other projects presented at the competition.

References

[1] I. T. Hjaltalín and H. Sigurdarson, "The strategic use of ai in the public sector: A public values analysis of national ai strategies," *Government Information Quarterly*, vol. 41, 02 2024.

[2] Y. Kumar and N. Chikkaguddaiah, *A Deep Learning Based System to Estimate Crowd and Detect Violence in Videos*, 09 2023, pp. 45–57.

[3] S. Das, *Real-time and Intelligent Security enablement through Edge AI-enabled CCTV Cameras : Santosh Das*, 04 2023.

[4] "Real-time violence detection in surveillance videos using deep learning approach," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, pp. 1267–1274, 04 2024.

[5] B. Sabiri, B. Asri, and M. Rhanoui, *Efficient Deep Neural Network Training Techniques for Overfitting Avoidance*, 07 2023, pp. 198–221.

[6] M. Pasupuleti, *Advancing Real-Time Intelligence with Edge AI*, 04 2024, pp. 94–110.

[7] F. U. M. Ullah, "A study of sequential deep learning- based methods for violence detection in videos," Ph.D. dissertation, 02 2023.

[8] K. Talha, K. Bandapadya, and M. Khan, "Violence detection using computer vision approaches," 06 2022, pp. 544–550.

[9] S. Sasmal, "Data engineering best practices with ai integration," 01 2024.

[10] V. Yadav, "Ai and human rights: A critical ethico-legal overview-international-review.com cc by nc 2023," vol. 14, pp. 261–270, 12 2023.

[11] D. Todaro, *Public Sector AI Applications in Shanghai*, 03 2024, pp. 295–554.

[12] M. Shubber and Z. T. Al-Taai, "A review on video violence detection approaches," *International Journal of Nonlinear Analysis and Applications*, vol. 13, no. 2, pp. 1117–1130, 2022.

[13] S. Lomlen, "Impact of artificial intelligence in enhancing national security," *Artificial Intelligence Studies*, vol. 1, 02 2024.

[14] H. Gupta and S. T. Ali, "Violence detection using deep learning techniques," in *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, 2022, pp. 121–124.

[15] K. Aarthy and A. A. Nithya, "Crowd violence detection in videos using deep learning architecture," in *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 2022, pp. 1–6.

[16] M. Cheng, K. Cai, and M. Li, "Rwf-2000: An open large scale video database for violence detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4183–4190.

[17] Y. Zuo, A. Hamrouni, H. Ghazzai, and Y. Massoud, "V3trans-crowd: A video-based visual transformer for crowd management monitoring," in *2023 IEEE International Conference on Smart Mobility (SM)*, 2023, pp. 154–159.

[18] S. Vosta and K.-C. Yow, "A cnn-rnn combined structure for real-world violence detection in surveillance cameras," *Applied Sciences*, vol. 12, no. 3, 2022.

[19] A. Chauhan and R. Gupta, "Human violence detection using lhogf algorithm and deep learning model," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2022, pp. 1202–1206.

[20] R. G. Tiwari, H. Maheshwari, A. K. Agarwal, and V. Jain, "Hybrid cnn-lstm model for automated violence detection and classification in surveillance systems," in *2023 12th International Conference on System Modeling and Advancement in Research Trends (SMART)*, 2023, pp. 169–175.

[21] N. Bagga, G. Singh, B. Balusamy, and A. Shanker Singh, "Violence detection in real life videos using convolutional neural network," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2022, pp. 872–876.

[22] Y. Lyu and Y. Yang, "Violence detection algorithm based on local spatio-temporal features and optical flow," in *2015 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration*, 2015, pp. 307–311.

[23] D. K., V. L.K.P., and C. S., "Autocorrelation of gradients based violence detection in surveillance videos," *ICT Express*, vol. 6, no. 3, pp. 155–159, 2020.

[24] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the airtlab dataset," *IEEE Access*, vol. 9, pp. 160 580–160 595, 2021.

[25] Y. Su, G. Lin, J. Zhu, and Q. Wu, "Human interaction learning on 3d skeleton point clouds for video violence recognition," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 74–90.

[26] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, "Efficient two-stream network for violence detection using separable convolutional lstm," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[27] G. Garcia-Cobo and J. C. SanMiguel, "Human skeletons and change detection for efficient violence detection in surveillance videos," *Computer Vision and Image Understanding*, vol. 233, p. 103739, 2023.

[28] H. Mohammadi and E. Nazerfard, "Video violence recognition and localization using a semi-supervised hard attention model," *Expert Systems with Applications*, vol. 212, p. 118791, 2023.

[29] R. Hachiuma, F. Sato, and T. Sekii, "Unified keypoint-based action recognition framework via structured keypoint pooling," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 962–22 971.

[30] Y. Qiao, W. Cui, and T. Shi, "Lam-2srn: A method which can enhance local features and detect moving objects for action recognition," *IEEE Access*, vol. 8, pp. 192 703–192 712, 2020.

[31] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," 06 2012, pp. 28–35.

[32] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," 06 2012, pp. 1–6.

[33] J. Nam, M. B. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, vol. 1, pp. 353–357 vol.1, 1998.

[34] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017.

[35] N. Habeeb, R. Thomas, and K. Oommen, "Automated detection of pneumonia using pre-trained convolutional neural networks in x-ray images," in *2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, 2023, pp. 1–6.

[36] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 4544–4554, 04 2016.

[37] J. Lin, J. Li, J. Gao, W. Ma, and Y. Liu, "Jointly modeling spatio-temporal features of tactile signals for action classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 13 817–13 825, 03 2024.

[38] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[39] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, and X. Zhang, "Tensorflow: A system for large-scale machine learning," 05 2016.

[40] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[41] V. Deepa, C. S. Kumar, and T. Cherian, "Ensemble of multi-stage deep convolutional neural networks for automated grading of diabetic retinopathy using image patches," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part B, pp. 6255–6265, 2022.

[42] Y. Ma, "Parallel programming: Driving the computational surge in ai," *Applied and Computational Engineering*, vol. 37, pp. 197–201, 02 2024.

[43] J. Yan, Y. Chen, and W. Fok, "Detection of children abuse by voice and audio classification by short-time fourier transform machine learning implemented on nvidia edge gpu device," 08 2023, pp. 893–897.

[44] S. Ji, Y. Tan, T. Saravirta, Z. Yang, Y. Liu, L. Vasankari, S. Pan, G. Long, and A. Walid, "Emerging trends in federated learning: from model fusion to federated x learning," *International Journal of Machine Learning and Cybernetics*, pp. 1–22, 04 2024.

[45] L. Patil, "From pixels to diagnoses: Exploring the potential of hyper-spectral imaging and artificial intelligence in plant disease detection," 05 2023.

VISION & MISSION OF THE DEPARTMENT

Vision

To achieve excellence in Artificial Intelligence and Data Science to cater to the ever-changing industrial and socio-economic needs.

Mission

- To provide high-quality and value-based technical education in the Artificial Intelligence and Data Science program.
- To establish an infrastructure fostering industry-institute interaction in order to meet global expectations and requirements.
- To empower students to become globally competent and effective problem-solvers to develop entrepreneurial skills and higher studies.