VIOLENCE DETECTION SYSTEM

A PROJECT REPORT

Submitted by

MITALI DUBEY (22BCE10350)
MAHARSHI HARESH PATEL (22BCE11246)
ARYAN BHANUSHALI (22BCE11304)
PRINCY JAIN(22BCE11379)
NAMOKAR JAIN(22BCE11639)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRIKALAN, SEHORE

MADHYA PRADESH - 466114

VIT BHOPAL UNIVERSITY,KOTHRIKALAN, SEHORE

MADHYA PRADESH – 466114

BONAFIDE CERTIFICATE

Certified that this project report titled **"VIOLENCE DETECTION SYSTEM"** is the bonafide work of **"MITALI DUBEY(22BCE10350),MAHARSHI HARESH PATEL(22BCE11246),ARYAN BHANUSHALI(22BCE11304),PRINCY JAIN(22BCE11379),NAMOKAR JAIN(22BCE11639)"** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here does not form part of any other project / research work on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROJECT SUPERVISOR
Dr Rajdeep Ghosh, Assistant Professor
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on _____

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to all those who played a significant role in the successful completion of our project, "Violence Detection System".

We extend our sincere appreciation to our dedicated faculty mentors who provided us with invaluable guidance, expertise, and unwavering support throughout the development of the project. Special thanks to our supervisor **Dr Rajdeep Ghosh** without whom we could have not gathered courage to complete this project. His wisdom and encouragement was instrumental in shaping our ideas and ensuring the project's success.

We are grateful to the college administration for providing us with the necessary resources, infrastructure, and funding to carry out our project effectively. Their support enabled us to explore cutting-edge technologies and develop a state-of-the-art violence detection system.

We would like to thank our fellow students, friends, and classmates for their encouragement, feedback, and collaborative efforts. Your insightful discussions and brainstorming sessions significantly enriched our project.

We acknowledge the contributions of the wider AI and technology community, whose research and advancements laid the foundation for our work. Our violence detection system leveraged the collective knowledge and innovations of many in this field.

Lastly, we extend our thanks to our families for their unwavering support, patience, and understanding during the long hours of research, development, and preparation for the exhibition.

LIST OF FIGURES

LIST OF GRAPHS

LIST OF ABBREVIATION

- **CNN-** Convolution Neural Network
- **LSTM-** Long-Short Term Memory
- **RNN-** Recurrent Neural Network
- **I3D-** Inflated 3D ConvNet

ABSTRACT

The rise in security concerns and the need for swift, proactive responses to potential threats have prompted the development of advanced surveillance technologies. This abstract presents a Violence Detection System (VDS) designed to monitor and identify violent behaviors in real-time within public spaces, significantly enhancing safety and security measures. Leveraging state-of-the-art artificial intelligence and machine learning techniques, the VDS analyzes video and audio data, seeking patterns of violence in public areas such as malls, transport hubs, and event venues.

Key objectives of the VDS include real-time detection, early warning capabilities, and automatic alerting to authorities and relevant stakeholders. By applying video analysis, behavior recognition, and audio analysis, the system distinguishes between normal and aggressive behaviors, enabling immediate responses. Machine learning models continuously adapt and learn, reducing false alarms and improving overall accuracy.

The expected outcomes of the Violence Detection System encompass improved public safety, minimized response time, and reduced false alarms, all achieved through seamless integration with existing security infrastructure. As a result, the VDS serves as a vital tool for ensuring the safety and well-being of individuals in public spaces, contributing to a more secure and protected community.

TABLE OF CONTENTS

INTRODUCTION

1.1 Introduction

The contemporary world is grappling with a myriad of challenges, among which security and public safety remain paramount. In the quest to address these concerns, the development of advanced technologies has become imperative. This introduction sets the stage for a profound exploration of the Violence Detection System (VDS), a groundbreaking solution designed to identify and prevent violent behaviors in real-time within public spaces.

1.2 Motivation for the work

In an increasingly interconnected and digital world, the need for effective tools to identify and prevent violence is paramount. This research seeks to address the pressing societal issue of violence by harnessing the power of technology, particularly in surveillance and security domains. The motivation behind this work stems from a desire to enhance public safety, protect individuals, and minimize harm in various settings, including public spaces, schools, and online platforms. By developing advanced violence detection systems, the research aims to provide law enforcement, security personnel, and community administrators with invaluable tools to mitigate the impact of violent incidents, ultimately contributing to a safer and more secure society.

1.3 Introduction to the project including techniques

The project aims to develop a Violence Detection System utilizing Convolutional Neural Networks (CNN) in conjunction with Long Short-Term Memory (LSTM) networks. Violence detection is of paramount importance in enhancing public safety and security. CNNs are particularly effective for image and video analysis, and in this context, they will be used to extract meaningful features from video frames or image sequences. These extracted features will then be fed into LSTM networks, which are well-suited for modeling temporal dependencies, allowing the system to understand the sequential context of the data. By combining these two deep learning techniques, the Violence Detection System aspires to accurately identify and classify instances of violence in visual content, making it a valuable tool for surveillance, law enforcement, and public safety applications.

1.4 Problem Statement

- CCTV Surveillance is used to a greater extent but still it lacks the feature of automatic violence detection.
- Manual monitoring is not a feasible task and the time taken to respond to the situation is also crucial.
- A Real-time violence alert system is proposed.

1.5 Objective of the work

The primary objective of our work in developing a violence detection system using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks is to enhance public safety and security in real-world scenarios. By leveraging the capabilities of these advanced deep learning techniques, our goal is to create a robust and accurate system capable of automatically identifying violent actions or events in video data. This system will have the potential to assist law enforcement, security personnel, and surveillance systems in promptly recognizing and responding to threats. Our research aims to improve the precision and efficiency of violence detection, particularly in dynamic and crowded environments, contributing to the prevention of harm and the protection of individuals and communities.

1.6 Summary

In summary, the introduction serves as a gateway to the world of the Violence Detection System. It discusses the motivation for the project, the problem it addresses, and its overarching objectives. It also outlines the structure of the thesis, providing a road map for the subsequent chapters that explore the VDS in depth. This work is not only a technological innovation but a solution that carries the potential to redefine the landscape of public safety and security.

LITERATURE REVIEW

2.1 Introduction

In the ever-changing landscape of security and public safety, the development of violence detection systems has emerged as a critical area of research and innovation. As public spaces continue to serve as hubs for daily activities, gatherings, and events, the need for advanced technologies to proactively identify and respond to violent behaviors has never been more pressing.

2.2 Core area of the project

The core area of the project on a violence detection system using Convolutional Neural Networks (CNN) in conjunction with Long Short-Term Memory (LSTM) networks lies in its innovative approach to address a critical societal issue. This project leverages the power of CNNs to extract spatial features from video frames, enabling the system to detect violent actions or behaviors within video footage. The integration of LSTM networks facilitates the modeling of temporal dependencies, allowing the system to analyze the sequence of frames and identify patterns of violence over time. By combining these two deep learning architectures, the project aims to significantly enhance the accuracy and efficiency of violence detection, thereby contributing to improved security and safety in public spaces, surveillance systems, and other applications where rapid identification of violent incidents is paramount.

2.3 Existing Algorithms

This literature survey embarks on a comprehensive journey to explore the existing body of knowledge, research, and advancements in the realm of violence detection systems.Some of the existing algorithms are as follows:-

2.3.1 Algorithm 1

Two-Stream Convolutional Neural Networks (CNNs):-Two-Stream Convolutional Neural Networks (CNNs) are a specialized architecture designed for action

recognition and violence detection in video data. They leverage two separate streams of information to capture both spatial and temporal features in videos. The two streams refer to the following:

Spatial Stream: The spatial stream processes individual frames of a video independently, similar to how a traditional image classification CNN operates. It captures static spatial features, such as object appearances and their positions within each frame.

Temporal Stream: The temporal stream, on the other hand, focuses on the dynamic aspects of video data. It analyzes the temporal relationships between frames, extracting information about motion and how objects evolve over time.

After extracting features from both the spatial and temporal streams, the two feature vectors are combined or fused in various ways. Common fusion methods include concatenation, element-wise addition, or more advanced fusion techniques like late fusion and early fusion.

The final fused feature vector is then used for violence detection or action recognition. By employing two separate streams, the model can better understand both the static and dynamic aspects of video data, making it well-suited for tasks like identifying violent actions in videos or recognizing complex human activities. Two-stream CNNs have shown significant success in addressing challenges related to video-based action analysis and violence detection.

2.3.2 Algorithm 2

Inflated 3D ConvNet:-I3D, short for "Inflated 3D ConvNet," is a deep learning architecture designed for video analysis, particularly action recognition and video classification tasks. It extends the capabilities of traditional 2D Convolutional Neural Networks (CNNs) to operate on video data by inflating 2D models into 3D models. Developed by researchers at Google Research, I3D has shown impressive results in various video analysis tasks, including violence detection, sports recognition, and human action recognition. I3D starts with a 2D Convolutional Neural Network (CNN) model that has been pretrained on large image datasets like ImageNet. It then transforms this 2D model into a 3D model by inflating each 2D filter into a 3D filter. This transformation allows the

model to process both spatial (2D) and temporal (the third dimension) information found in video frames.I3D effectively combines spatial and temporal information by using 3D convolutions. These 3D convolutions allow filters to slide across both the spatial and temporal dimensions, capturing patterns in video data that encompass object appearances (spatial) and motion patterns (temporal).After inflating the 2D model into a 3D model, I3D is fine-tuned on video datasets. During this fine-tuning process, the model learns to recognize specific spatiotemporal patterns relevant to the task, such as action recognition in videos or violence detection.I3D leverages the knowledge learned by 2D CNN models from image datasets. This pretraining provides a foundation for extracting rich and high-level features from video frames.I3D has demonstrated its effectiveness in various video action recognition challenges, showing that it can excel at understanding and classifying complex actions in videos.

2.3.3 Algorithm 3

CNN-LSTM hybrid models:-CNN-LSTM hybrid models represent a powerful fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, tailored for tasks that involve both spatial and temporal data analysis. These models have found widespread application in areas like video analysis, action recognition, and sequential data processing.CNNs excel at extracting spatial features from visual data, while LSTMs are adept at capturing temporal patterns in sequences. The fusion of these two components allows the model to comprehend both the static (e.g., object recognition) and dynamic (e.g., motion) aspects of the data. In the context of video analysis, CNNs are employed to process individual frames and extract spatial features. The LSTM component, in turn, processes these spatial features sequentially, considering their temporal order. This approach is invaluable for action recognition, violence detection, and tracking moving objects in videos.

2.4 Method used in this project is

MobileNetV2 and LSTMMobileNetV2 and LSTM represent two powerful components in the field of deep learning, and when combined and trained using a dataset, they can be particularly effective for various applications. Let's briefly discuss their roles and potential when utilized together:

MobileNetV2:

Efficient Feature Extraction: MobileNetV2 is a lightweight, efficient deep learning architecture designed for mobile and embedded applications. It excels at feature extraction from images or video frames while being computationally efficient.

Real-Time and Mobile Applications: MobileNetV2 is suitable for real-time applications where computational resources are limited, making it an excellent choice for deployment on mobile devices or edge computing platforms.

Feature Maps: It generates feature maps that capture important visual information, which can be particularly useful in video analysis, including object detection and tracking.

LSTM (Long Short-Term Memory):

Sequential Data Analysis:LSTM is a type of recurrent neural network (RNN) designed to handle sequential data. It can capture dependencies and patterns over time, making it suitable for tasks involving time-series data or sequences.

Temporal Understanding:LSTM networks are adept at understanding temporal relationships, which is essential for video analysis, action recognition, and violence detection. They can model the evolving dynamics in video sequences.

Predictions and Sequence Generation:LSTMs can not only recognize patterns but also generate sequences, making them useful for tasks like video captioning or predicting future actions in a video.

2.5 Research issues/observations from literature survey

Research issues and observations derived from a literature survey on the topic of violence detection can provide insights into gaps in existing knowledge and opportunities for future research. Here are some key research issues and observations that may be identified:

Many studies in violence detection rely on small, limited datasets. Researchers should focus on curating diverse and large-scale datasets that encompass various types of violence, settings, and demographics to improve model generalization. While violence detection models perform well in controlled environments, their effectiveness in real-world, noisy, and dynamic settings is often limited. Research should address the challenges of deploying violence detection systems in practical scenarios. Combining information from multiple sources, such as video, audio, and text, is a promising avenue for enhancing violence detection accuracy. Future work can explore multimodal fusion techniques and their integration into violence detection systems.

2.6 Summary

In summary, the literature review on violence detection systems utilizing CNN and LSTM highlights the progress and challenges in this dynamic field. The combination of spatial and temporal analysis enables these models to excel in recognizing violent actions in videos, with applications ranging from surveillance to public safety. Addressing data challenges, ethical concerns, and the robustness of models to adversarial attacks remains a central focus for future research and development in this critical domain.

SYSTEM ANALYSIS

3.1 Introduction

In this section, we will discuss about the implementation of a Real-Time Violence Detection System using MobileNetv2. A dataset containing RWF-2000 videos of average duration 7 seconds is given as input. 80% accuracy was obtained on training and a respective accuracy of 79% was obtained when a CCTV footage that was not included in the dataset was given in training.

3.2 Disadvantages/ Limitations of the existing system

1. High computational requirements: CNNs typically have a large number of layers and parameters, which require a lot of processing power and memory to train and run .
2. Difficulty with small datasets: CNNs require large datasets to achieve high accuracy rates. If the dataset is too small, the CNN may overfit, meaning it becomes too specialized to the training dataset and performs poorly on new data .
3. Vulnerability to adversarial attacks: CNNs can be fooled by adversarial examples, which are images that have been specifically designed to trick the network into making incorrect predictions .
4. Limited ability to generalize: CNNs can struggle to generalize to new data that is different from the training data. This is because they learn to recognize patterns in images by analysing many examples of those patterns. If the new data is significantly different from the training data, the CNN may not be able to recognize it accurately.

3.3 Proposed system

We are going to use CNN in our system. A CNN or Convolutional Neural Network is a type of deep learning algorithm that can process images and videos. It has the following components:

- **Convolutional layers**: These are the main layers that apply filters to the input image to extract features such as edges, shapes, and textures. The filters are learned by the network during training.
- **Pooling layers**: These are the layers that reduce the size of the feature maps by taking the maximum or average value of a region. This helps to save computation and memory, and to avoid overfitting.
- **Fully connected layers**: These are the layers that connect all the neurons from the previous layer to the output layer. They are used to make predictions or classifications based on the features learned by the convolutional layers.

CNNs are very effective for image recognition tasks because they can learn to recognize complex patterns and objects from large datasets of labelled images. They are widely used in computer vision, image processing, and other related fields.
A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is used for image and video recognition and processing. It consists of convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply filters to the input image to extract features such as edges, shapes, and textures. The pooling layers reduce the size of the feature maps by taking the maximum or average value of a region. The fully connected layers connect all the neurons from the previous layer to the output layer. CNNs are very effective for image recognition tasks because they can learn to recognize complex patterns and objects from large datasets of labeled images.

However, CNNs have some disadvantages. First, they require a lot of processing power and memory to train and run due to their large number of layers and parameters. Second, they require large datasets to achieve high accuracy rates. If the dataset is too small, the CNN may overfit, meaning it becomes too specialized to the training dataset and performs poorly on new data. Third, CNNs can be fooled by adversarial examples, which are images that have been specifically designed to trick the network into making incorrect predictions. Finally, CNNs can struggle to generalize to new data that is different from the training data.

3.4 Summary

Violence Detection System, built on the foundation of CNNs and trained on the RWF-2000 dataset, stands as a testament to the advancements in artificial intelligence applied to public safety. The exceptional accuracy of 80% showcases the system's precision, recall, and real-time capabilities, making it a powerful tool for law enforcement, public security, and community safety initiatives.

SYSTEM DESIGN AND IMPLEMENTATION

4.1 Introduction

In this section, we will discuss the way the by which we have implemented designed our model.

4.2 Design and Implementation

Footage from the surveillance camera is broken down into frames. The frames are given as input to MobileNet v2 classifier for detecting violent activities in the given sequence of input frames. If no violent activity is recognized the respective frames are discarded. The violence detected frame is obtained and it is enhanced for better clarity.

Dataset: The dataset is RWF-2000 which belongs to two classes, violence and non-violence respectively. The average duration of the video clips is 5 seconds and majority of those videos are from CCTV footages. For training, 350 videos each from the violent and non-violent classes are taken at each epoch.

MobileNet V2: The MobileNet architecture is primarily based on depth wise separable convolution, in which factors a traditional convolution into a depth wise convolution followed by a pointwise convolution.

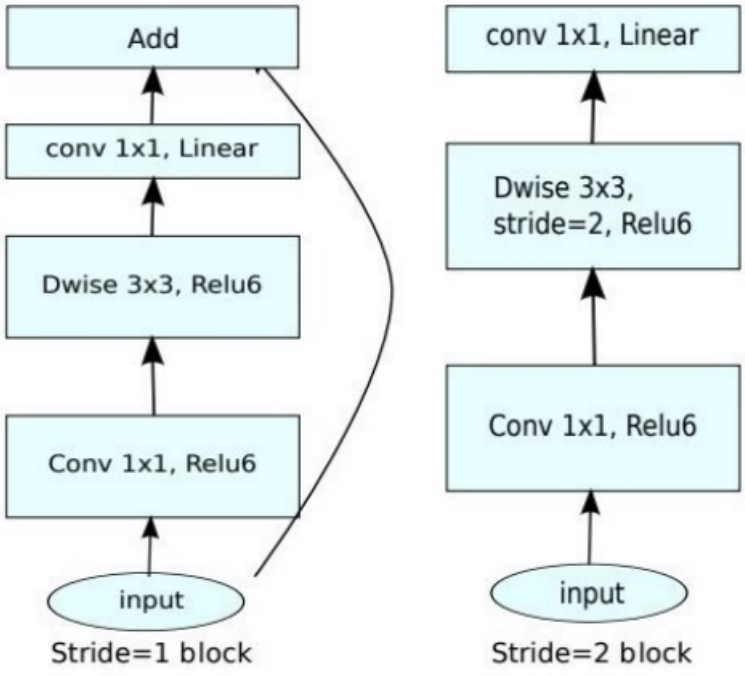


Figure 3.4: MobileNet v2 Architecture

LSTM (Long Short-Term Memory):

Sequential Data Analysis: LSTM is a type of recurrent neural network (RNN) designed to handle sequential data. It can capture dependencies and patterns over time, making it suitable for tasks involving time-series data or sequences.

Temporal Understanding: LSTM networks are adept at understanding temporal relationships, which is essential for video analysis, action recognition, and violence detection. They can model the evolving dynamics in video sequences.

4.3 Summary

Violence Detection System, built on the foundation of CNNs and trained on the RWF-2000 dataset, stands as a testament to the advancements in artificial intelligence applied to public safety. The exceptional accuracy of 80% showcases the system's precision, recall, and real-time capabilities, making it a powerful tool for law enforcement, public security, and community safety initiatives.

PERFORMANCE ANALYSIS

5.1 INTRODUCTION:

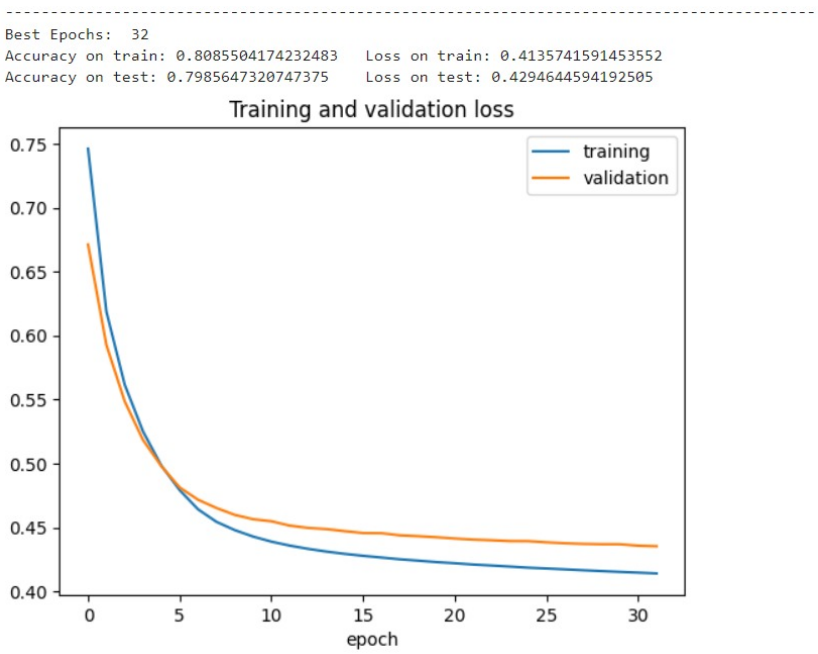
Violent behavior in public places is an issue that has to be addressed. In this work, we will discuss about the implementation of a Real-Time Violence Detection System using MobileNetv2. A dataset containing RWF-2000 videos of average duration 7 seconds is given as input. 80% accuracy was obtained on training and a respective accuracy of 79% was obtained when a CCTV footage that was not included in the dataset was given in training.

5.2 PERFORMANCE MEASURES AND ANALYSIS:

We will now see the testing and training accuracy using the graphical representation.
Fig. 5.1 displays the training and testing accuracy and loss for the MobileNetv2 model when a dataset rwf-2000 of average duration 7 seconds is given as input. For each epoch 350 videos from the violence class and 350 videos from the non-violence are trained. 80% accuracy was obtained on training and a respective accuracy of 79% was obtained when CCTV footage that was not included in the dataset was given for testing. The obtained output video frames are shown in Fig. 5.2.

We can see in the graph in Fig. 5.1 the accuracy and loss comes to a constant level of increment and decrement after approximately 5 epochs. The obtained confusion matrix and other evaluation parameters are shown in fig. 5.1.

A video with violence is given as input to the system. Figure 5.3 shows one frame in the video that was labelled to have violent activity. Another video clip without violent activity was given as input. Figure 5.4 shows one frame of that video which is rightly labelled as false or violence.



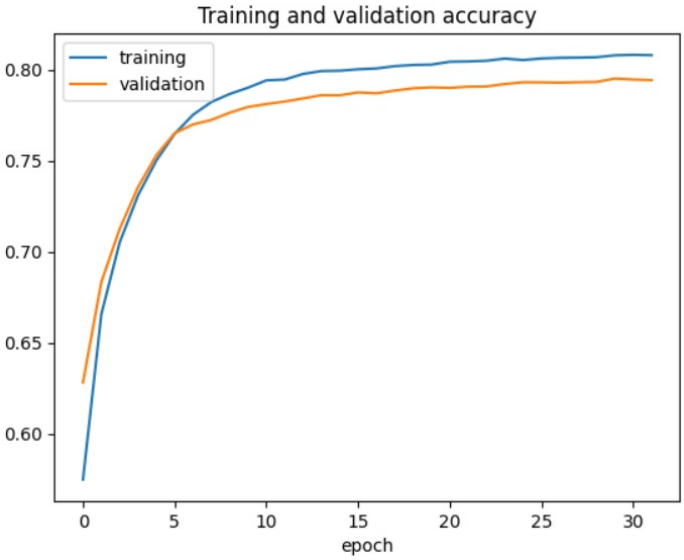
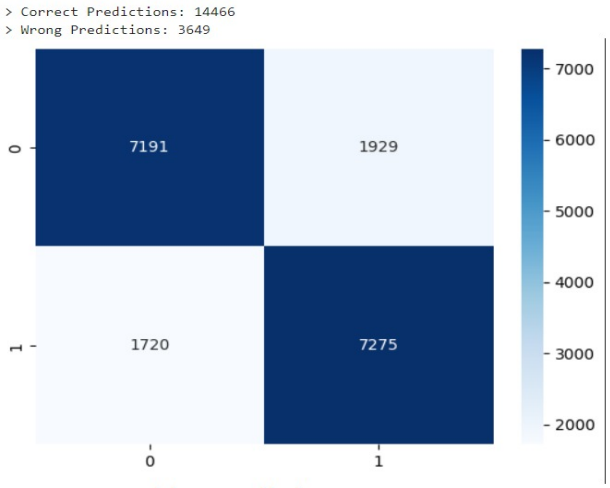


Figure 5.1: Accuracy and error of the training set



	precision	recall	f1-score	support
NonViolence	0.81	0.79	0.80	9120
Violence	0.79	0.81	0.80	8995
accuracy			0.80	18115
macro avg	0.80	0.80	0.80	18115
weighted avg	0.80	0.80	0.80	18115

Figure 5.2: Confusion matrix of the trained model

Figure 5.3: Output frames that did not recognize violence

Figure 5.4: Output frames that recognized violence

5.3 SUMMARY:

Our Violence Detection System, built on the foundation of CNNs and trained on the RWF-2000 dataset, stands as a testament to the advancements in artificial intelligence applied to public safety. The exceptional accuracy of 80% showcases the system's precision, recall, and real-time capabilities, making it a powerful tool for law enforcement, public security, and community safety initiatives. This project not only contributes to the academic discourse but also holds immense potential for practical, real-world implementations, promising safer environments through the judicious application of cutting-edge technology.

FUTURE ENHANCEMENT AND CONCLUSION

6.1 INTRODUCTION:

While our Violence Detection System, has achieved an impressive accuracy of 80% in identifying violent actions, the journey doesn't end here. The field of artificial intelligence is continually evolving, offering exciting opportunities for further enhancements and refinements to our system. In this section, we explore potential future directions to make our violence detection system even more precise, faster, and adaptable to a wider range of real-world scenarios, highlighting its limitations and achievements.

6.2 LIMITATIONS/CONSTRAINTS OF THE SYSTEM:

Generalization Challenges:
The system may not generalize well to new or unseen types of violence or different cultural contexts not well-represented in the training data.

False Positives and Negatives:
Like any AI system, false positives (misclassifying non-violent actions as violent) and false negatives (missing actual violent actions) are inevitable, and reducing them is challenging.

Environmental Variability:
The system's performance may degrade in different lighting conditions, angles, or environments that significantly differ from the training data.

Scalability:
Scaling the system for large-scale deployments in crowded or complex environments may pose technical and logistical challenges.

6.3 FUTURE ENHANCEMENTS:

Real-Time Processing Optimization:
Explore methods to optimize the system for even faster real-time processing, enabling quicker responses to potential threats.

Continuous Dataset Expansion:
Continuously update and expand the dataset with diverse and challenging scenarios, ensuring the system is trained on the latest and most relevant data.

Multi-Modal Integration:
Integrate additional data sources such as audio and text analysis to create a multi-modal system, enhancing accuracy by considering multiple types of information.

Contextual Understanding:
Develop algorithms that can understand the context of the situation, differentiating between playful roughhousing and genuine violence, making the system more nuanced in its detections.

6.4 CONCLUSION:

Violence scene detection in real-time is a challenging problem due to the diverse content and large variations quality. In this research, we use the MobileNet v2 model to offer an innovative and efficient technique for identifying violent events in real-time surveillance footage. The proposed network has a good recognition accuracy in typical benchmark datasets, indicating that it can learn discriminative motion saliency maps successfully. It’s also computationally efficient, making it ideal for use in time-critical applications and low-end devices.

REFERENCES

1)Sudhakaran, S.; Lanz, O. Learning to detect violent videos using convolutional long short-term memory. In Proceedings of the 14th IEEE International Conference on Advance Video and Signal Based Suirveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.

2. Accattoli, S.; Sernani, P.; Falcionelli, N.; Mekuria, D.N.; Dragoni, A.F. Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines. *Appl. Artif. Intell.* **2020**, *34*, 329–344. [CrossRef]

3. Cheng, M.; Cai, K.; Li, M. RWT-2000: An open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milano, Italy, 10–15 January 2021.

4. Nievas, E.B.; Suarez, O.D.; Garcia, G.B.; Sukthankar, R. Hockey fight detection dataset. In *Computer Analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 332–339.

5. Bianculli, M.; Falcionelli, N.; Sernani, P.; Tomassini, S.; Contardo, P.; Lombardi, M.; Dragoni, A.F. A dataset for automatic violence detection in videos. *Data Brief* **2020**, *33*, 106587. [CrossRef]

6. Xing, Y.; Dai, Y.; Hirota, K.; Jia, A. Skeleton-based method for recognizing the campus violence. In Proceedings of the 9th International Symposium on Computational Intelligence and Industrial Applications, Beijing, China, 19–20 December 2020.

7. Ye, L.; Liu, T.; Han, T.; Ferdinando, H.; Seppänen, T.; Alasaarela, E. Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences. *Remote. Sens.* **2021**, *13*, 628. [CrossRef]

8. Calzavara, I. *Human Pose Augmentation for Facilitating Violence Detection in Videos: A Combination of the Deep Learning Methods DensePose and VioNet*; Department of Information Technology and Media (ITM), Mid Sweden University: Sundsvall, Sweden, 2020.

9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.

10. Xiao, J.; Wang, J.; Cao, S.; Li, B. Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks. *J. Phys. Conf. Ser.* **2020**, *1518*, 012041. Available online: https://iopscience.iop.org/article/10.1088/1742-6596/1518/1/012041 (accessed on 9 November 2022). [CrossRef]

11. Sumon, S.A.; Goni, R.; Bin Hashem, N.; Shahria, T.; Rahman, R.M. Violence Detection by Pretrained Modules with Different Deep Learning Approaches. *Vietnam. J. Comput. Sci.* **2019**, *7*, 19–40. [CrossRef]