National College of Ireland

Enhancing Surveillance Security Through Violence Detection Using Advanced Deep Learning Algorithms

MSc Research Project
Masters in Data Analytics

Ashish Negi
Student ID: 23237198

School of Computing
National College of Ireland

Supervisor: Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ashish Negi
Student ID:	23237198
Programme:	Masters in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Aaloka Anant
Submission Due Date:	12/12/2024
Project Title:	Enhancing Surveillance Security Through Violence Detection Using Advanced Deep Learning Algorithms
Word Count:	6979
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	23rd January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Surveillance Security Through Violence Detection Using Advanced Deep Learning Algorithms

Ashish Negi
23237198

Abstract

Detection of violent activities is a prime importance in terms of public safety, security monitoring, and law enforcement support. The growing dependence on extensive surveillance systems in the public and private domains made it necessary to ascertain these violent acts in real-time, which is quite a challenge. Violence detection in surveillance videos becomes a critical task with public safety, law enforcement, and security monitoring applications. Though quite challenging, real-time detection of violent activities remains difficult to accomplish due to the dynamic nature of video data, constraints on computational efficiency, and the need to be accurate across diverse situations. Existing solutions mostly rely on traditional and special techniques of computer vision or single deep learning models, which can get bogged down while performing both tasks of higher computation efficiency and accuracy in a complex environment. This paper presents a comprehensive framework that harnesses advanced deep learning algorithms: Dense Neural Networks, Long Short Term Memory, Gated Recurrent Units, and a hybrid LSTM+GRU model, for the task. Our methodology combines spatial and sequential feature extraction from video frames, preprocessing, data augmentation, and model training. Evaluation of these models is performed using accuracy, precision, recall, F1-score, AUC, and loss to identify the best model. The GRU model outperformed all, achieving slightly better accuracy and generalization, making it the best possible solution for any real-life application. As a practical application, we have developed a Flask-based web application so that users can upload videos, which could lead to detecting violent activities.

1 Introduction

The advancement of technology in video-making and image processing is considered to be significant in providing useful content in terms of meeting various applications. These applications are identified with recognizable actions and objects, which are often used by attackers, thus introducing a path to violence. The visual representation of this violence such as CCTV footage or videos in surveillance cameras alternatively has facilitated helpful progress for law enforcement agencies in visually monitoring the environment and protecting society. Violence has a major impact on society, which is continuously increasing amid digital advances. The potential implication of this information entails both positive and negative aspects of a technologically-enhanced society. A major focus on violence detection has continuously evolved with specific actions and activities by analyzing video datasets, indicating unethical human actions that are indicated as violence.

Traditionally, violence detection has been carried out manually based on observing violent acts directed through an individual or a group.

The footage inspection based on a manual procedure has been identified as a tedious and not-so-feasible approach. Dealing with this violence, particularly physical assaults has marked a new research paradigm of detection. The reciprocated knowledge from the literature has emphasized various long-term, mid-term, and short-term detection solutions. Against the downturn in security and surveillance approaches to effectively recognize violence, several state-of-the-art methods have been introduced that have presented critical tactics to tackle the complex problems typically raised by violent crimes. A seamless integration of advanced technologies such as deep learning algorithms has presented an architecture, that has been recognized as a shift in contemporary research. The potential approach to detecting the damage caused by violent crimes has enabled the incorporation of more proactive measures although the sophisticacy of these methods needs further investigation to measure the accuracy of the detection procedure. One of the potential reasons for leveraging deep learning models in the current research study is to determine the potentiality of the architecture based on video feeds (datasets). Moreover, it would enable enforcement officials to enhance the investigation process of violence by identifying and documenting the critical situation.

1.1 Research Objectives

- To design an efficient deep-learning framework for detecting violence in surveillance videos.
- To compare Dense, LSTM, GRU, and Hybrid models using key performance metrics to identify the best approach.
- To develop a Flask-based web application for violence detection in uploaded videos.

1.2 Research Question

- How can advanced deep learning models, including Dense Neural Networks, LSTM, GRU, and Hybrid architectures, be effectively designed, evaluated, and implemented to detect violence in surveillance videos, and how can this framework be extended to develop a scalable web application for real-world and real-time violence detection?

2 Related Work

This chapter has provided critical evidence from multiple studies considerably focused on violence detection and the relevance of video feeds. Over the years, research has systematically presented evidence on how the violence detection procedure is mandatory to recognize potential crime events harming society. In this focus, different solutions have been introduced by experts and demonstrated by researchers in their studies, thus presenting the academic relevance of these proactive measures. Hence, the chapter contextualizes specific events of violence and further introduces measures that can effectively address such circumstances.

2.1 Violence Detection Using Traditional Methods

The immense difficulties experienced over the years with violent actions have presented comprehensive insights into how violent crimes, particularly physical assaults have increased in recent times. According to the information provided by Chandane et al. (2024) effective ways to address such issues have introduced optimized solutions to detect real-time violence based on footage and video feeds. Traditionally, law enforcement teams along with crime detection experts manually handle the situations. With the investigation of local methods, the evidence presented by Gao et al. (2016) explained that the spatio-temporal approach to feature points has been identified as a novel technique, which was traditionally used by experts to address the detection constraint from the manual detection process. As represented by the study, the arrangement has focused on a "bag of words" approach that intends to explore possibilities based on learning from the probable distribution of feature points. Despite a suitable implication of the approach, limited interesting points to gather meaningful information from the crime situation can be established. In another study presented by Zhang et al. (2016) the information on violence detection has presented significant insights into the surveillance system, which is focused on action recognition through vision-based methods. The study shows the specific limitation of the method in recognizing violence, especially in determining the location of the crime.

With the drawback of the visual method, Zhang et al. (2016) introduced another solution based on the "Gaussian Model of Optical Flow" (GMOF) that has been proposed to extract candidate-based violence regions, thus leveraging modeled adaptive insights compared to traditional normal crowd behavior determined from crime scenes. The experimental outcome achieved from the study shows that the detection accuracy of the new model is higher in terms of speed processing even in immense crowded scenes. In recent decades where big data has been playing a vital role in introducing a smart living system, concerns have been directed to usage perspectives such as enhanced surveillance that contributes to huge data generation. Although the novelty of this data has been defined by many studies, the computation, and precision of these enormous data is a challenge, making it difficult to apply in recognition activities Mumtaz et al. (2023). Understandably, the research paradigm based on violence detection has presented facts entailing how the detection of anomalous actions is challenged by manually devised engineered features. The overview of the information presented by Mumtaz et al. (2023) comprehensively dives deeper into producing insights into potential dilemmas faced with the reliance on traditionally engineered features. The Real-time image processing and introduction of improved datasets have explained the necessity of domain advancement to improved methods that can provide in-depth analysis of feature extraction from relevant datasets and highly accurate detection of violence.

A critical awareness presented with the demonstration of intelligent solutions for video surveillance and effective image processing has recognized the need for revolutionary trends in violence detection. Yao and Hu (2023) explained that analyzing anomaly behavior from an aggressive scene, for example, physical assaults can be a useful dataset to positively detection and restrain violence. Amid this understanding, the above study emphasized a future trend that demonstrates basic principles of "Violent Behaviour Detection" (VioBD) based on the comparison and evaluation of performed datasets. Thus,

the literature evidence obtained from existing studies has identified a need to introduce enhanced detection processes including detection methods and datasets.

2.2 Violence Detection Using Machine Learning Methods

It is aware that violence, precisely physical assaults is a common concern of today's society. Over the past few decades, there have been extensive studies on this particular area based on which various detection techniques are introduced. According to the information provided by Ramzan et al. (2019) violence detection based on surveillance videos has been identified as a common method in the detection process. Understandably, several other techniques are also developed simultaneously to detect critical violence such as brutal events and harmful patterns often identified in videos. In this regard, certain parameters like acceleration, time-specific, flow, and appearance are typically used to collect attributes (features) from videos to detect violent activities. Ramzan et al. (2019) in their study have explained the purpose of computer vision as a suitable method of violence detection by analyzing surveillance cameras. Abnormal behavior and violent actions are common areas in this kind of research where computer vision and image processing using state-of-the-art methods have provided significant outcomes. As per the study findings, traditional ML models such as support vector machines (SVM) have been used to analyze features from video datasets, which play a vital role in violence recognition. The experimental outcome determined from the approach shows a helpful contribution although future work is required in the focused domain.

Surveillance systems have become an efficient method for city safety and protection task adherence for law enforcement agencies. With its importance in the detection of violence and abnormal behaviors through automatic recognition of video-based scenarios, surveillance systems are commonly used in multiple areas such as railway platforms, schools, banks, and hospitals. One of the potential implications of this surveillance system-based automatic violence detection is to quick recognition of unethical actions, thus helping law enforcers to take immediate measures. Upon understanding the progress, Biswas et al. (2022) explained that the introduction of machine learning algorithms, for example, support vector machines (SVM) and hybrid models have provided significant information into feature extraction and violence detection based on video datasets. Comparatively, another study presented by Omarov et al. (2022) has investigated analysis methods for violence detection and thus intends to disassemble existing conditions. In this focus, the study precisely demonstrated emerging trends in violence discovery through continuous research and investigation. A suitable emphasis on state-of-the-art methods has addressed specific problems in existing techniques in video-based violence detection, especially with manually-engineered datasets. Omarov et al. (2022) provide comprehensive insights into video-based violence detection and consider approximately 154 pre-existing studies to identify, screen, and analyze information to understand the credibility of ML methods. The findings gathered based on the review insights show that methodologies that have considered ML methods and end-to-end DL methods provide successive results with higher detection accuracy with public datasets.

The above information has been summarised to understand the series of progression in violence detection constituted for years from manual handling to automatic recognition. However, a future tendency in the detection process is an additive focus of researchers.

A study presented by Ullah et al. (2023) has emphasized the automatic violence detection mechanism through intelligent classification solutions, for example, machine learning (ML) and "artificial neural networks" (ANN). The study explained that extremely overcrowded areas, likely, subways, public places, and banks are in dire need of violence detection (VD) systems for safety assurance and security. The demonstration presented by Ullah et al. (2023) on VD applications has highlighted the working mechanism of ML techniques and the importance of surveillance datasets to obtain an accurate detection outcome. However, the above study has marked the specificity of datasets and the concurrent challenge faced by researchers in using video surveillance datasets. Understandably, the research contribution has presented ultimate guidelines indicating recommended attributes to present trending research in the arena with suitable datasets that can benefit the detection process. The focus and emphasis on violence detection methods have been explored through the lens of the rising assaults in recent years - a global concern that has severely affected society.

In the study presented by Singh et al. (2018) the evidence has determined cases of street harassment, which is observed at the rate of 79% among Indian women, 86% among Thailand women, and 89% among Brazil women. On the other hand, it has been further identified that in London, nearly 75% of women faced physical assaults or public violence. Hence, ensuring safety is necessarily deemed through practical measures fundamental to the protection of society. Upon understanding the priority, Singh et al. (2018) have explained that with several limitations in existing methods in terms of poor detection accuracy, high false-alarm rates, and high computation costs, the introduction of modern methods such as machine learning techniques provides evolutionary insights. Findings established from the study indicate that violence-related features when applied through motion tracking segregate video frames as per the presence of certain moving objects. While understanding the flow of events, the calculation of this optimal flow is a critical step for each video frame. The approved "violent flow descriptors" have been further applied with several ML methods and comparatively evaluated. The findings further explained that the implementation of feasible techniques through fusion has been applied in the practice, which has improved both detection accuracy and reduction of error rates.

The discussion of research practices that have been presented by researchers based on the detection of behavioral anomalies has imparted knowledge on novel techniques. In the contemporary era, with the penetration of the Internet, higher data transmission has been identified. While this evolution has become a key factor in prediction research, especially violent prediction, preventing children from witnessing TV programs that contain violent scenes is important. According to the information provided by Khan et al. (2019) violence is a common phenomenon that has been widely accessed in real life and movie shows. While prioritizing societal safety, the physical and mental health of individuals is an essential consideration to prevent adverse impacts. Thus, violence detection has become a vital research focus. While understanding this, Khan et al. (2019) specified the proposition of novel transfer learning algorithms that have been tested using benchmarked violence datasets. Considerably, the experimental outcome achieved shows a fast, feasible, and accurate mechanism of violence detection. This information indicates that research orientation on the continuous improvement in detection methods has alleviated knowledge of how improved algorithms can provide better outcomes than traditional state-of-the-art methods.

2.3 Violence Detection Using Deep Learning Methods

Previously, it has been identified that the complexities in datasets and poor accuracy outcomes with increased error rates in existing detection techniques have continuously presented the need for emerging intelligent solutions. Although machine learning methods have provided suitable outcomes in detecting violence and abnormal behavior, the reliability of the technique is affected by the datasets' complexities. According to the information presented by Sumon et al. (2020) it has been explained that salient features of datasets comprising violent and non-violent images and videos from different events provide significant insights into training and testing models. While machine learning models show limitations to some extent, the contribution of deep architectures has presented successive outcomes. In the above study, the importance of convolutional neural network (CNN) and long-short-term memory (LSTM) models have been explained in terms of feature extraction and trained using a movie-based violent dataset. Suggestively, the experimental outcome obtained shows an accuracy level of 97.6% in violent detection. The potential implication of the study is to determine how these improved models have shown feasibility in the detection process irrespective of the dataset pattern or configurations.

It is aware that physical aggression has become a serious problem for society and subsequently impacts people worldwide. The impact of these events therefore draws attention to exploring the root cause and ensuring safety across high-crime areas. In this regard, the information provided by Negre et al. (2024) explains that real-time detection of violence and physical assaults using advanced algorithms such as deep learning models can provide efficient solutions to address the need for human supervision and people safety. The study has presented insights into AI-based and deep neural architecture integrated video analysis for detecting violence. Indicating the detection process, it has been identified that deep architectures can provide potential solutions in using video datasets for efficient violence detection. The supporting evidence presented in another study by Arun Akash (2022) has demonstrated a combined approach of computer vision and using inception models such as v3 and Yolo-v5 to detect violent actions. The study emphasized the contribution of deep models in the detection process by evaluating their efficiency to utilize videos from crime scenes. With its "application programming interface" (API) ability to real-time detection of violence, an accuracy level of 74% has been obtained. The estimation shows that there is a greater possibility of detection improvement using deep learning models.

The subjective knowledge established through empirical evidence on violence, especially physical assaults has informed the need to tackle these threats with improved measures. In this regard, widespread deployment of "video surveillance" has been identified as a major solution applied by law enforcers for years for visually monitoring the environment and implementing prompt actions. Understandably, Fatima Kiani and Kayani (2022), explained that utilizing CCTV footage as live streaming datasets for real-time detection using deep learning models can diminish the error rate that often occurs in human supervision. The study inferred knowledge of the efficiency of the solution and marked it as a viable solution to assess the issue. Fatima Kiani and Kayani (2022), have demonstrated the contribution of LSTM and Yolo inception models to promote real-time detection of video objects and further classification. At the same time, the study has also introduced DeepSort as an improved algorithm that suitably augmented the efficiency of the

above-combined models. While DeepSort has augmented the efficiency of the ensemble model, the application of relevant datasets for training and testing the model has further enhanced the model interface. Indicating the outcome of the proposed technique, it has been identified that a full-fledged detection system has been introduced by leveraging deep architectures that have increased detection efficiency with enhanced tests and evaluations.

Empirical evidence gathered from current studies has specifically drawn attention to the efficiency of deep learning models in achieving higher accuracy in violence detection. According to the information presented by Soliman et al. (2019) automatic recognition/detection of violence based on videos constituting mass people or crowds is a broader researcher interest. The key deliverables from this detection emphasize the need to prevent violence and protect society from harm. Understandably, the introduction of deep learning or neural architecture in practice has served the purpose with ease, thus gaining enormous attention in the context. The above study develops an understanding of the importance of the LSTM model as a "temporal feature extractor" consisting of sequences based on fully connected layers, thus facilitating the classification process. The approach is significant and achieved an accuracy of 88.2% based on violence and non-violence videos as datasets. Comparatively, another study presented by Wang et al. (2021) has presented insights into the purpose of utilizing video surveillance systems as a basis for intelligent classification and detection of brute forces harming society. This study has further prioritized face recognition, which is imperative to enhance controlled measures. In this regard, the application of artificial features that have in-depth properties serves a purpose in extracting spatiotemporal features from videos using a CNN network. The model combines further with trajectory features.

The above-specified SPP-enhanced CNN model has been tested using video datasets containing violent features typically established from 'Crow' and 'Hockey' events. Indicting the experimental outcome, it has been identified that for each case the accuracy achieved is approximately 92% and 97.6%. Indicating the experimental outcome, it can be stated that over the years, the integration of DL models with improved dataset arrangement has provided improved outcomes. Kaur and Singh (2022) explained that both CNN and LSTM-based deep architecture are continuously reviewed in the literature to explore the reliability of both techniques. Based on the information from previous studies, it has been identified that both models are efficient in violence detection although CNN performs much better than the other model. It is already noted from the extension review of the evidence that the dataset is a factor that influences a test result in violence detection. Sernani et al. (2021) introduce the AIRTLab dataset, which is a novel approach designed to facilitate the detection process. One of the potential implications of the dataset is to increase the robustness of applied deep architectures in addressing false positive alarm rates. Thus, it can be stated that research evidence not only emphasized the improvement in model applications but also prioritized novel datasets introduced from public repositories to improve the process of violence detection.

3 Methodology

Violence detection from surveillance CCTV camera videos is a challenging problem that has implications for public safety and surveillance. A model has to be extremely robust

and accurate to differentiate between violent and non-violent actions. The connectivity across temporal and spatial patterns within very long video sequences poses a major challenge, giving rise to misclassification and compromising reliability. In this section, we will discuss in detail all the steps that will be incorporated into our method from data pre-processing to feature extraction, followed by model training and evaluation. The methodology diagram for detecting the violence activities from surveillance videos is shown is shown in Figure 1.

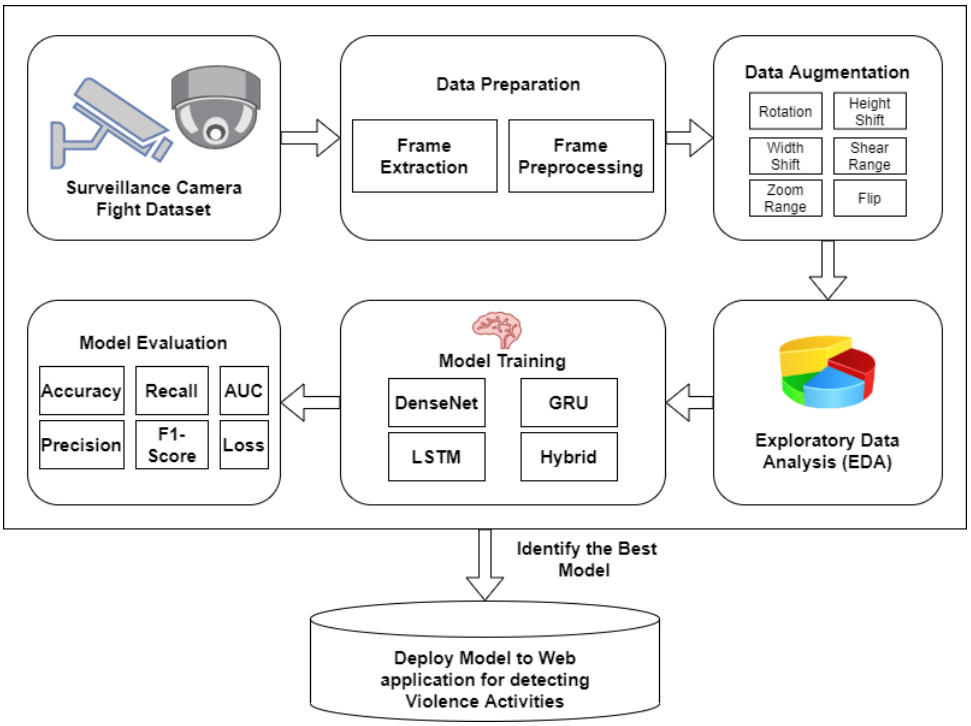

Figure 1: Methodology Framework for Detection of Violence Activities

3.1 Dataset Description

The Surveillance Camera Fight Dataset is a well-structured collection aimed at creating systems to detect fights in public surveillance environments like streets, public transport stations, and other public venues *GitHub - seymanurakti/fight-detection-surv-dataset: New generated dataset for fight detection in surveillance cameras. — github.com* (n.d.). It consists of 300 videos equally balanced between fight and non-fight instances lasting 2 seconds. The data includes clips of fights in uncontrolled situations, such as hitting with objects, kicking, punching, and wrestling, in diverse environments including streets, cafes, and buses. The videos were collected from YouTube to ensure high relevance by enforcing an opportunity for learning in contexts where background motion is minimal, thus making it suitable for surveillance applications. Also, it provides a balanced dataset to be used for binary classification tasks, which in turn enables us to effectually train and evaluate the models in real-time fight detection from video surveillance.

3.2 Data Preparation

The data preparation process involves specifying the input and output paths for the datasets of violence and non-violence video streams and their respective extracted frames. The frame extraction from the corresponding videos is achieved via a set of evenly spaced indices that ensure complete passage of the video duration. Five frames are extracted as the minimum and a maximum of ten frames per video, and these are organized in two folders, one for fight and the other for noFight. Sample plots are created to visualize the extracted frames for quality checks as shown in Figure 2.

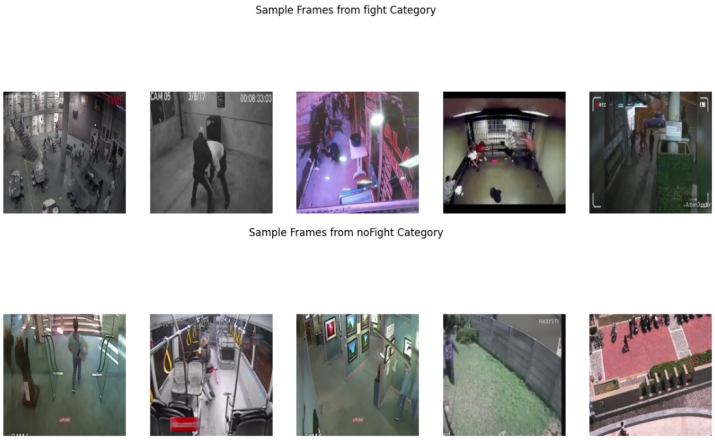

Figure 2: Sample Frames for Fight and No-Fight category

After extracting the frames, they are fed into pre-processing, where they are resized to a standard dimension of 224x224 pixels, and their pixel values are normalized between 0 and 1. The processed frames are saved into a distinct folder according to their respective labels. Finally, the dataset is stored in memory, with each frame labeled as either 1 for fight or 0 for noFight. The dataset is then split into training set (70%), validation (20%), and test set (10%) modes, ensuring an equal representation of both categories.

3.3 Data Augmentation and Visualization

To improve model generalization and safeguard against overfitting data augmentation techniques are applied to the training dataset. Augmentations consist of random rotations, width and height shifts, shear transformations, zooming, and horizontal flips. These augmentations generate diverse variations of the original training images as shown in Figure 3.

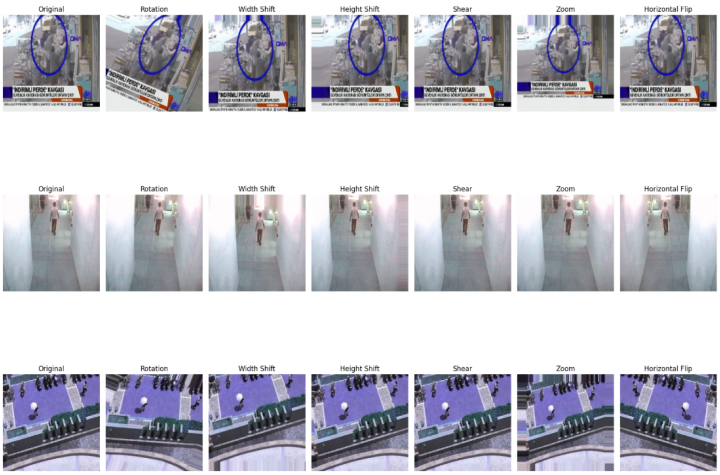

Figure 3: Sample Frames after data augmentation

Separate data generators are created for the training, validation, and testing data sets, so augmentation is only used with the training data. Thereafter, some random samples from the training set are visualized together with their augmented counterparts to elucidate the effect of the augmentation techniques. Through this process, the dataset is enlarged, allowing the model to undergo a well-trained fit without compromising on validation and testing data rigor.

3.4 Exploratory Data Analysis

The bar chart in Figure 4 illustrates the distribution of frames within the categories of fight and noFights, demonstrating nearly balanced categories, with 1500 frames each. The resulting balance created a fair training procedure without allowing a bias to arise in favor of one category.

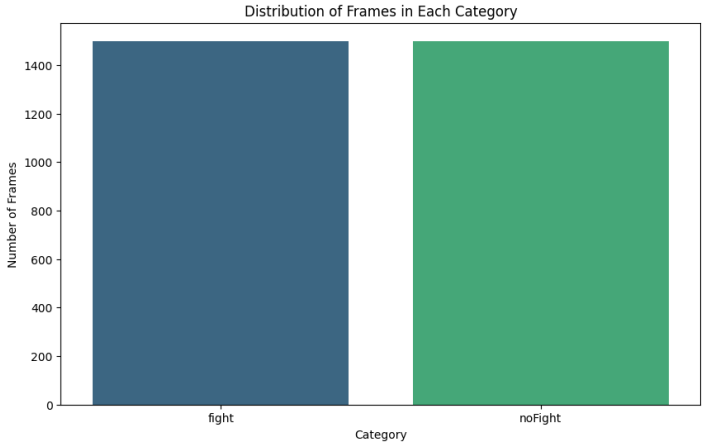

Figure 4: Frame Distribution Across fight and noFight Categories

The histograms in figure 5 characterize the distribution of average color intensities (Blue, Green, and Red channels) for the fight and noFight categories. These two categories identify overlapping yet fairly different patterns because the noFight frames distinctly

show the highest intensity in the blue channel. This information shows the visual differences in color features that can help the model distinguish between the two categories.

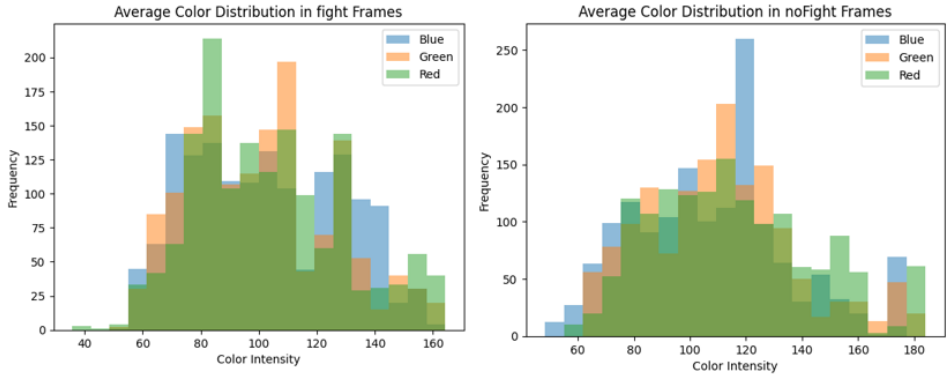

Figure 5: Average Color Intensity Distribution in fight and noFight Frames

The histogram in Figure 6 reveals the brightness distribution of frames in the fight and noFight categories. Fight frames are generally relatively darker, clustered around 80-100, while noFight frames have a broader distribution that centers on higher brightness values. This shows that brightness could be a distinguishing attribute that might assist the model in the two categories' discrimination.

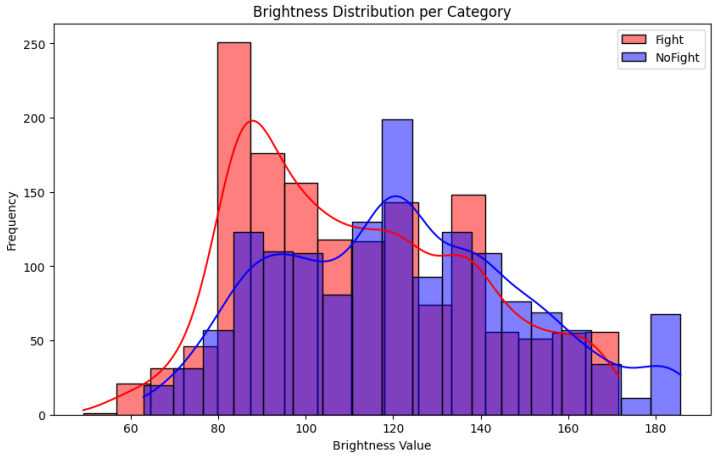

Figure 6: Brightness Distribution Across fight and noFight Categories

The Structural Similarity Index Measure (SSIM) is a measure of similarity between any two images which captures the differences due to structural changes, luminance, and contrast. The histogram in Figure 7 shows that the noFight frames have High SSIM values near 1.0, indicating strong frame-to-frame similarity and the fight frames express SSIM differently due to rapid movement and structural differences. The result suggests that SSIM can highlight differences in dynamics between both categories.

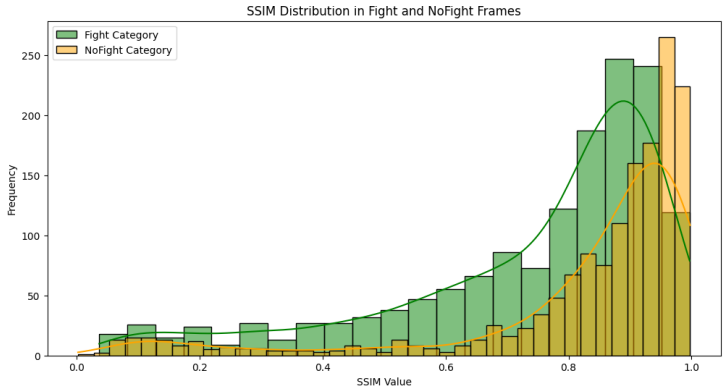

Figure 7: SSIM Distribution Between Consecutive Frames in fight and noFight Categories

The line graph in Figure 8 shows the brightness values for each frame in the fight and noFight categories. Brighter fight frames are notably distinctive with a greater variation in brightness owing to the movements, while noFight frames tend to have constant brightness.

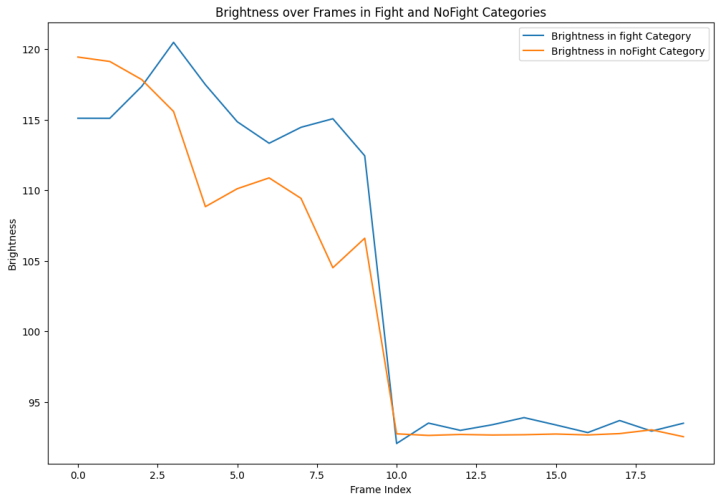

Figure 8: Brightness Variation Over Frames in fight and noFight Categories

3.5 Model Training

During model training, features are extracted from the preprocessed frames using a pre-trained MobileNetV2 Convolutional Neural Network (CNN) model. This model performs the required spatial encoding to produce high-dimensional feature vectors. The dataset is then divided into training, validation, and test sets, while the application of ImageDataGenerator for data augmentation introduces variations such as rotations, shifts, and flips to the training data. Multiple models, including a Dense Neural Network (DNN), Long Short-Term Memory (LSTM) network, Gated Recurrent Unit (GRU) network, and a Hybrid LSTM+GRU model, are defined to handle both spatial and sequential modes of behavior found in the data. After the features have been extracted, training is performed on these models, which are monitored over multiple epochs to optimize performance.

3.6 Model Evaluation

Model evaluation is an important step in the machine learning pipeline to ensure the reliability, robustness, and generalizability of trained models. During the model evaluation phase, various trained models are analyzed under accuracy, precision, recall, F1-score, AUC, and loss in the test dataset. These metrics are carefully chosen because they combine to give a good measure of how well a model is performing with regard to accuracy, balance between false positives and negatives, and the ability to classify overall. Bar charts compare these metrics against one another across different models, in order to underscore their advantages and disadvantages. The training and validation accuracy and loss versus epochs are plotted in line plots, offering insights into convergence, generalization, and tendencies toward overfitting.

We performed validation by splitting the dataset into three subsets: training (70%), validation (20%), and testing (10%). The validation set consisted of preprocessed frames of the videos, ensuring a balanced distribution of fight and non-fight classes. During training, the validation set was used to evaluate the model performance on unseen data after each epoch. This process helped fine-tune hyperparameters, like learning rate, batch size, and dropout rates. Metrics like validation accuracy, loss, and AUC were monitored to assess the model's generalization ability and detect overfitting or underfitting trends.

4 Design Specification

The architecture of an algorithm has a strong bearing on its ability to process the data with maximum efficiency in order to provide accurate predictions. Architectural understanding is crucial for choosing the right model for a given task and for improving its performance. This section describes the structural and architectural features of each algorithm used in this study.

4.1 Dense Neural Network (DNN)

A Dense Neural Network takes over high-dimensional visual feature vectors for the MobileNetV2 CNN. It detects the presence of violence in each frame of the video, based on simple patterns like shapes, edges, and textures, which may manifest in raised arms or a sudden movement. In a way, for each analyzed surveillance camera's footage, the DNN operates on spatial features computed within a single frame, which is useful in detecting violent acts in cases with low motion or when a single frame of view is being analyzed. Its straightforward architecture, with layer, support for dropping out those elements responsible for threat or disturbance with stable learning, overcomes static learning and, hence, is effective as a baseline for violence detection. The architecture of Dense Neural network is shown in Figure 9.

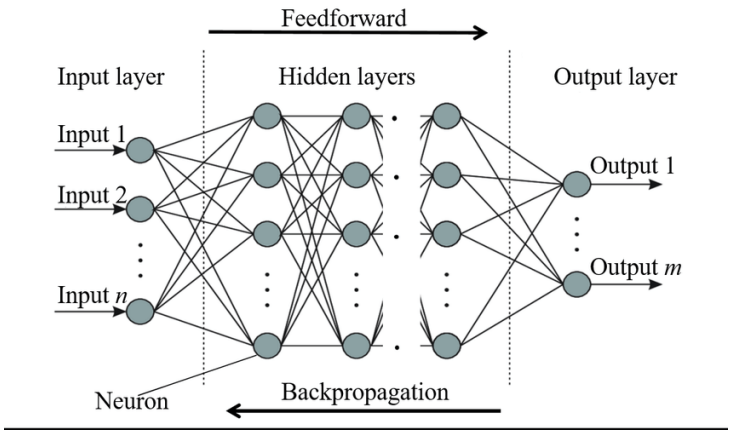

Figure 9: Dense Neural Network (DNN) Architecture Abueidda et al. (2022)

4.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory is a Recurrent Neural Network used for learning dependencies in sequential data, meaning it uses forget, input, and output gates in the context of selectively remembering. In violence detection of surveillance camera videos, the LSTM helps extract the patterns of time sequence of an escalation of a conflict or sudden, violent actions. It is particularly beneficial in analyzing continuous motions in video streams, as it can track and link sequences of movements across frames. Both short-range and long-range sequence property modeling by LSTM helps identify subtle temporal cues for violent behavior monitoring. The architecture of Long short-term memory is shown in Figure 10.

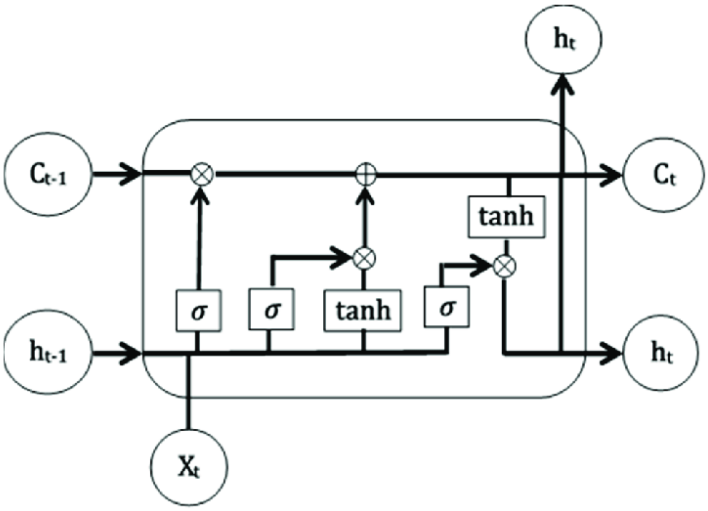

Figure 10: Architecture of Long Short Term Memory (LSTM) Toma et al. (2019)

4.3 Gated Recurrent Unit (GRU)

GRU is a simplified version of RNN, using reset and update gates to satisfy temporal dependence in a computationally efficient manner. Due to its advantages of saving time

and effort, GRUs are best suited for short video analysis in violence detection and work in situations without ample resources, like edge-based surveillance systems. These rapidly detect motion or some significant change of posture that looks like an initiation of violent action between consecutive frames. GRUs are further efficient for use in real-time with surveillance applications, thus setting up quick responses toward any detected violence. The architecture of Gated recurrent unit is shown in Figure 11.

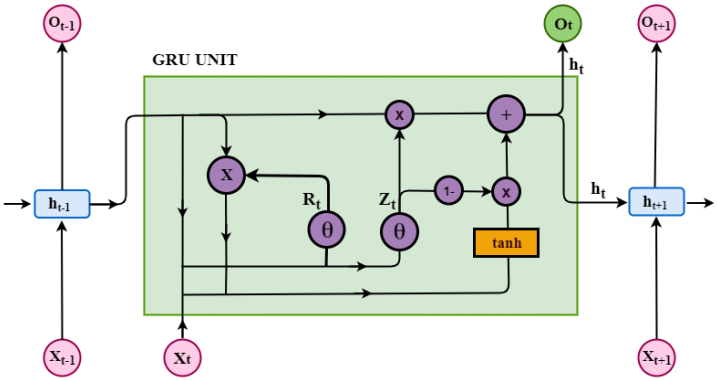

Figure 11: Architecture of Gated Recurrent Unit (GRU) Bibi et al. (2020)

4.4 Hybrid Model (GRU+LSTM)

The hybrid LSTM-GRU model provides a strong model offering a solution for violence detection by combining the long-term dependent learning of LSTM with the efficiency of GRU and being particularly suited for video clips taken from surveillance cameras in which prolonged as well as rapid actions need to be scrutinized. The LSTM layer captures the time-series display of events from an escalation of aggression, while the GRU layer finds sudden spikes in motion that are indicative of violence. Such an integrated approach guarantees reliable detection in dynamic and complex environments, such as crowds or very frantic environments. The architecture of hybrid model is shown in Figure 12.

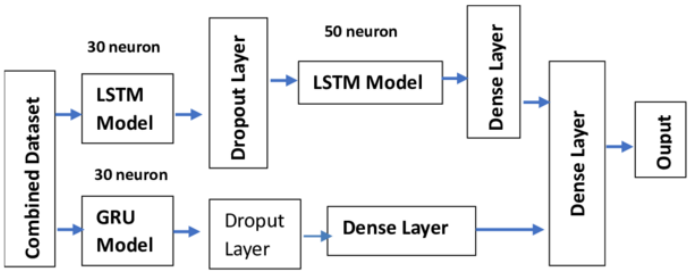

Figure 12: Hybrid Model (GRU+LSTM) Architecture Girsang and Stanley (2023)

5 Implementation

The implementation of a violence detection system used different Python libraries, with each having a specific role to play in the project. The os and cv2 libraries were employed

for file path management and video frame extraction, pickle for saving and loading extracted features enabling efficient reuse, and random and numpy to bring randomness into sampling and reproducibility. Libraries such as math and matplotlib.pyplot enabled calculations and visualization of data while Seaborn and plotly.express were useful for detailed exploratory data analysis (EDA) plots for the distribution of frames, brightness, and color attributes. For dataset preparation, pandas aided in structured manipulation of the data, and sklearn.model_selection was used to stratify and split the dataset into training, validation, and test sets. For preprocessing and augmentation of the video frames, tensorFlow.keras.preprocessing.image offered utilities for resizing, normalizing, and applying augmentations of rotation and flipping. Some of the pre-trained models utilized for feature extraction were MobileNetV2, ResNet50, and EfficientNetB0 from tensorflow.keras.applications, and their preprocessing utilities ensured they were ready for use directly into the architecture selected. The feature extraction task was performed via spatially encoding video frames into high-dimensional vectors using the pre-trained CNNs. To build, train, and test custom models, tensorflow.keras offered building blocks typically used for creating a sequential architecture such as Dense Neural Networks (DNN), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), and Hybrid models mixing LSTM and GRU layers. Some of the layers used were Dense, LSTM, GRU, BatchNormalization, Dropout, and Reshape with the use of EarlyStopping and ModelCheckpoint as callbacks for the optimization of training. The performance of each model was evaluated by accuracy, precision, recall, F1-score, and AUC computed from sklearn.metrics.

The web application that we have developed using Flask is a practical use case of our violence detection system. It allows users to upload videos through a very simple user interface to analyze the uploaded video for violent activities with the help of pre-trained AI models. The home screen reveals the "Violence Detection System" banner and gives a simple file upload interface through the "Upload and Analyze" button that controls the analysis. The interface is targeted to be user-friendly for easy interaction and highlights the implementation scope of our research toward practical purposes. This application is a demonstration of how AI models could indeed enhance safety and security by analyzing CCTV footage in real-time. Currently, the web application is deployed in the local system. The screenshot of web application Detecting violence and non-violence activities is shown in Figure 13 and Figure 14.

The web application is currently deployed in the local environment and is mentioned in the report. After running the application, it can be accessed at http://127.0.0.1:5000. The application is built using Python Flask. However, as part of the future work, this application can be deployed in a cloud environment to enable broader accessibility and real-time usage.

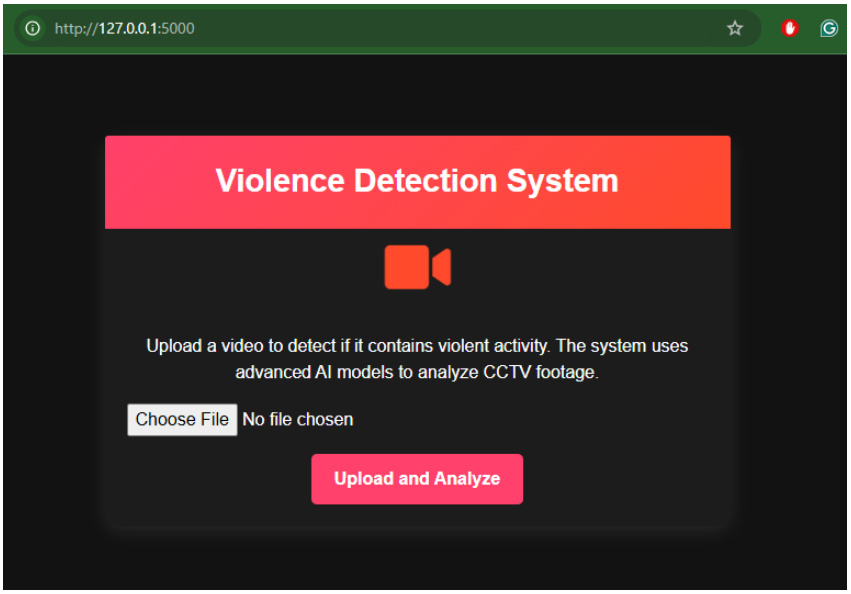

Figure 13: Web application interface for uploading videos for violence detection

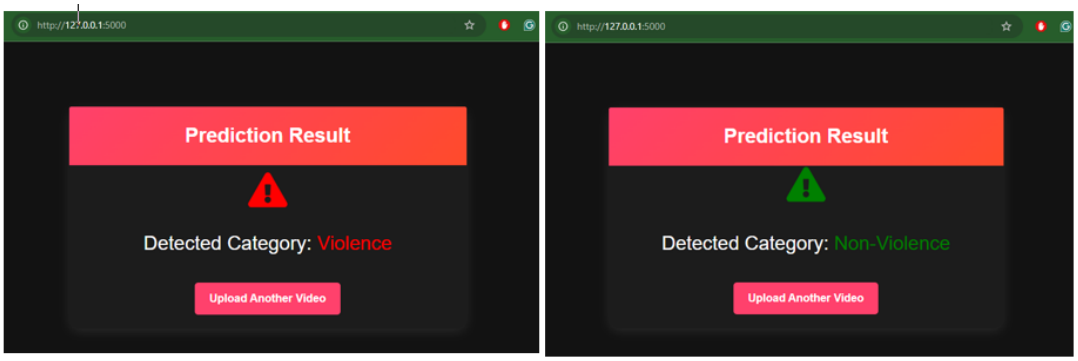

Figure 14: Prediction results of the web application showing detected categories Violence and Non-Violence

6 Evaluation

The evaluation of the model is significant as it helps to understand the models' strengths and weaknesses, shows the credibility of the models, and aids in selecting the best model for real-world deployment. In the evaluation section, we will examine the proposed models based on various metrics, namely accuracy, precision, recall, F1-score, AUC, and loss. This will provide an overall evaluation of how effective each model was in detecting violence in surveillance videos.

6.1 Experiment-1 / Accuracy Comparison

Accuracy is termed as the ratio of correctly predicted instances to the total instances in a data set, measuring the general performance of a model. It is a straightforward metric for the evaluation of classification tasks. Here, a high value of accuracy is desired

because it indicates that the model is able to predict correctly more often. The stacked bar chart in Figure 15 compares the accuracy between the performances of four models: GRU, Hybrid (LSTM + GRU), LSTM, and Dense. The Gated Recurrent Unit (GRU) was significantly outperformed and received the greatest accuracy, which was 96%, thus showcasing its proficiency in capturing temporal dependencies in video data. The hybrid model combining LSTM and GRU with accuracy at 89% stands next since it is good in representing spatial and temporal features. The LSTM and Dense models also had an equivalent performance with a throw of 66% on the accuracy test, which is significantly lower than that compared to GRU and Hybrid. This comparison shows how well the GRU is suited for temporal sequence modeling and how the hybrid model can span its advantages of combined architectures to boost efficiency in violence detection tasks.

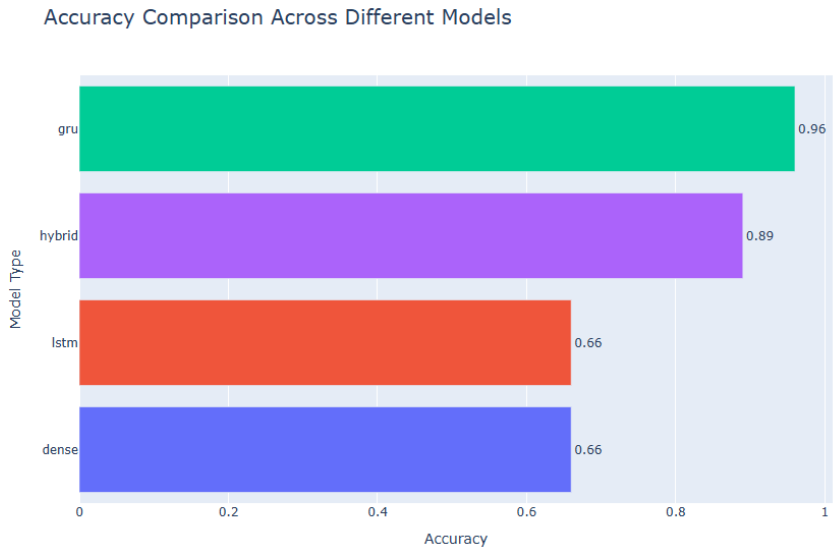

Figure 15: Accuracy comparison of different models used for violence detection

6.2 Experiment-2 / Precision, Recall and F1-Score Comparison

The precision measures the ratio of accurately predicted positive cases over all predicted positives, thus giving the measure of correctness of the model to classify positive instances. Recall, or Sensitivity, is the ratio of correctly predicted positive cases to the actual number of positive cases. It shows how well the model is capable of capturing true positives. The F1 score is defined by the harmonic mean of precision and recall. It balances the trade-off between precision and recall. Higher values of the three above indicate better model performance in general, while special consideration to the F1 score while dealing with imbalanced datasets is important to guard against any one of Precision or Recall vying for dominance. The bar chart in Figure 16 shows the comparison of Precision, Recall, and F1-Score of the Dense, LSTM, GRU, and Hybrid models indicating their performance in violence detection. The GRU model surpassed all the others, achieving perfection in all aspects with the scoreboard on Precision, Recall, and F1-Score was 1.0, which means that it is regarded as the best one in violence detection with little to no false positives and negative predictions. The Hybrid model followed close by with Precision at 1.0, Recall at 0.91, and F1-Score at 0.91, displaying a strong capability of moderation in detection accuracy. Dense and LSTM have also scored 1.0 on Precision, while their Recall and

F1-Scores have dropped down to about 0.66 and 0.65, respectively, indicating their poor performance in recognizing actual violent activity. This analysis demonstrates that the GRU model is the most effective in detecting violence, while the Hybrid model is the best alternative, and Dense and LSTM models are rather poor in this task.

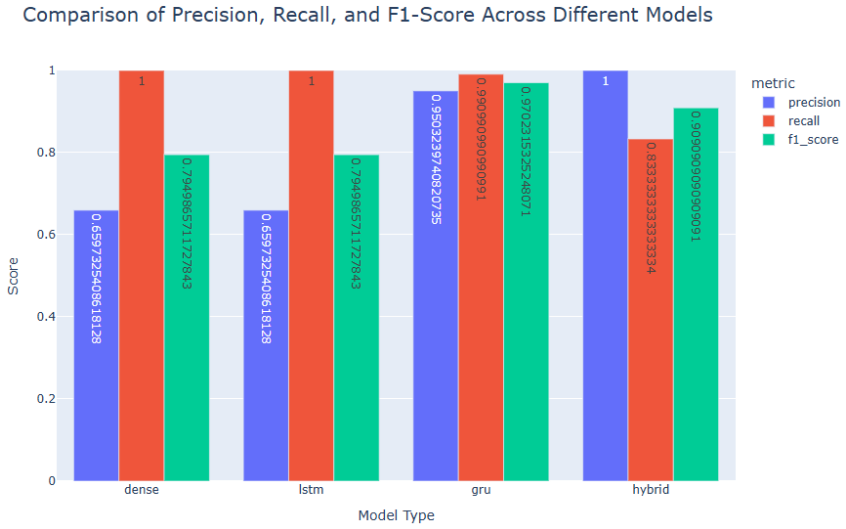

Figure 16: Precision, Recall and F1-Score comparison of different models used for violence detection

6.3 Experiment-3 / AUC Score Comparison

The Area Under the Curve score is one of the most salient evaluation criteria applied in assessing the performance of any classification model, especially for binary classification tasks. It measures the discriminative ability of the model between the classes across all threshold values. AUC represents how well the model can separate the classes, with a higher AUC score being more favorable: The closer the scores toward 1, the better the classification model's performance, while an AUC score close to 0.5 suggests guesses. The horizontal bar chart in Figure 17 represents the AUC score plots for Dense, LSTM, GRU, Hybrid in the context of their capability to distinguish between violent and non-violent acts. The GRU model achieved the highest AUC score of 0.95; this shows better accuracy of correct classification with little overlap between the classes. The hybrid model closely followed it, registering an AUC score of 0.92, an indication of good discriminatory power too. The Dense and LSTM models scored in AUC lower than a mere unimaginable 0.50, meaning they are not better performers at discriminative power and predict near-random guessing at best. This analysis affirms that the GRU model distinctively brings about a clear distinction between violence and non-violence, with the hybrid model being a credible second.

Figure 17: AUC Score comparison of different models used for violence detection

6.4 Experiment-4 / Loss Comparison

The loss is a measure of the difference between model predictions and actual target values during training. It measures how well the model is learning. The lower the loss, the better the model performs because it is capable of predicting values that are closer to the actual ones. A reduction in loss over epochs means improved accuracy and generalization for the model. The line chart shows in Figure 18 the training and validation loss trends of each model (Dense, LSTM, GRU, and Hybrid) across 10 epochs. Lower loss values are better since they imply that the model fits the data well. The GRU and Hybrid models have a stable and minimal loss throughout the training process, thus showing a better convergence and generalization ability. On the contrary, the LSTM model shows wider fluctuations in the validation loss, especially after the first few epochs, indicating possible instability. The Dense model shows a relatively high validation loss, indicating it does not generalize well. The GRU model demonstrates a low total loss and, hence, is the most suited for the task.

Initially, the model's generalization capability was poor, as observed through a large gap between training and validation accuracy. Thus, to improve the generalization, various data augmentation techniques such as random rotations, shifts, zooming, and flipping were employed to create more diverse training samples. Additional regularization techniques such as dropout and batch normalization were used in the aim of preventing the model from overfitting to the training data. These measures helped the model learn generalized features, resulting in reduced overfitting and better performance on both validation and unseen test data.

We generated diverse variations of images using data augmentation to enhance the model's ability to generalize to unseen data. By applying specific transformations, such as rotation, shifting, zooming, and flipping, we simulated several real-world conditions, including variation in lighting, camera angles, and backgrounds, thus allowing the model to tackle divergent scenarios in surveillance environments with reduced overfitting during

training.

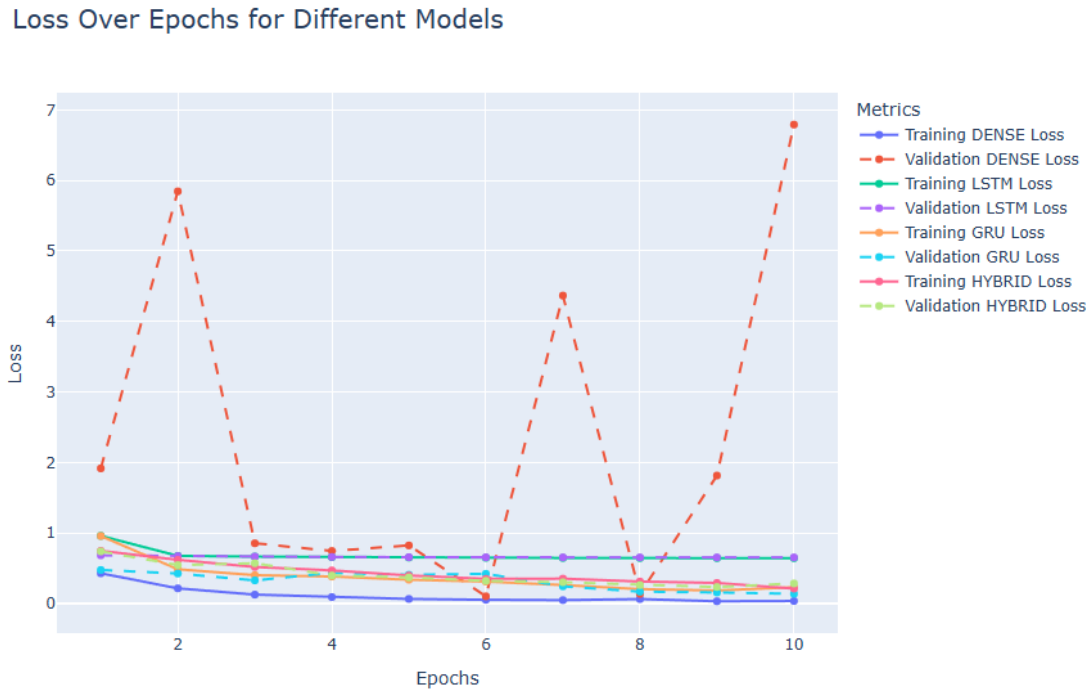

Figure 18: Loss over epochs comparison of different models used for violence detection

6.5 Discussion

This section provides a critical analysis of the experimental data and evaluation of the performance of the algorithms proposed in this study for violence detection in surveillance videos. The models were evaluated primarily on metrics such as accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and loss, combining the unique results obtained from four different models: Dense Neural Network, Long Short-Term Memory, Gated Recurrent Unit, and Hybrid LSTM-GRU. Our findings reveal that GRU invariably produces the best results in terms of accuracy (96%) and AUC (0.95), largely attributed to its efficiency in handling sequential dependency without experiencing the vanishing gradient problem common in RNNs. The Hybrid model combining LSTM and GRU achieved good results with slightly lower scores but showed more robustness through all metrics, as it utilized strengths from both architectures to handle temporal features. In LSTMs, the slightly sub-par performance (accuracy and AUC around 66%) can be attributed to their increased computational complexity and extended training duration, which may have resulted in over-fitting. With Dense Neural Networks, results fell as expected within a nearly uniform performance range with equally poor accuracy and AUC of 66% and 0.50 respectively, because of being ill-equipped to process video data temporal sequences. A vital aspect of the methodology relates to data preprocessing and feature extraction. Though MobileNetV2 worked well for spatial feature extraction, more modeling choices are available for exploring a better representation of features that include EfficientNet or ResNet50. Furthermore, although data augmentations ensured variability in training, their design could use improvements to reflect real-world scenarios more accurately, such as low-light conditions with camera distortion. Relative to previous studies, our findings are congruent with studies highlighting the ability of GRUs to

handle temporal data proficiently. Thus, in summary, GRU performed with remarkable results, owing to its capability in the analysis of sequential data, while the Hybrid model provided solid performance by composing different architectures. LSTMs encountered computational hurdles, and Dense models failed to learn temporal dependencies successfully.

We used 10 epochs because our experiments showed that the models converged within this range, with no significant improvement in performance beyond 10 epochs. This approach, computationally efficient and maintain the level of accuracy, limiting unnecessary resource utilization with maximum results.

7 Conclusion and Future Work

In this research, we have successfully developed a framework for violence detection in surveillance videos using deep learning techniques. The objective was to identify the best-performing architecture for this task by evaluating several models Dense Neural Networks, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and a Hybrid LSTM-GRU model. We successfully built a complete pipeline composed of preprocessing and feature extraction with MobileNetV2, model training, and evaluation. Furthermore, we developed a web application built using Flask that takes videos from users and utilizes the GRU model to predict violence, demonstrating the utility of our approach in practice. The GRU model is the best-performing model with the best accuracy of 96% and AUC of 0.95 because of its advantages in handling sequential dependencies. The hybrid model performed well, while LSTM and Dense performances suffered from computational complexity and lack of temporal processing capabilities, respectively. While our methodology showed promising results, there are certain limitations associated with it. The dataset used for this research is used for purposes of experimentation, and may not capture the full range of real-world scenarios such as changing lighting conditions, angles from which cameras capture videos, and occlusions. In addition, this version of the application permits offline analysis of uploaded videos and may not be useful for real-time surveillance systems.

In this work, we employed the NVIDIA A100 GPU from Google Colab for deep learning model training on violence detection. This particular infrastructure was not included in the NCI offer and was instead available on a paid subscription basis (Colab Pro), which I funded myself. The A100 GPU was essential to handle the computational demands of training complex models like GRU, LSTM, and Hybrid architectures on large datasets, enabling efficient processing and faster experimentation.

Future work will focus on extending the current implementation to support real-time violence detection in live video streams and as an integration into the security camera system for proactive monitoring. Attention mechanisms, like transformers, will further boost the model's attention to critical temporal features. Combining the GRU and Hybrid models by means of the ensemble approach may better robustness. The project also has high prospects for commercialization, especially in the security and surveillance domain, where automated violence detection systems find their sacred role in attending to public safety needs. These aspects must be duly addressed for this research to be further

developed into a workable and scalable real-time violence detection solution.

References

Abueidda, D., Koric, S., Abu Al-Rub, R., Parrott, C., James, K. and Sobh, N. (2022). A deep learning energy method for hyperelasticity and viscoelasticity.

Arun Akash, S. A., S. S. M. R. E. K. . N. N. (2022). Human violence detection using deep learning techniques, *Journal of Physics: Conference Series* **2318**(1).

Bibi, I., Akhunzada, A., Malik, J., Iqbal, J., Musaddiq, A. and Kim, S. (2020). A dynamic dl-driven architecture to combat sophisticated android malware, *IEEE Access* **PP**: 1–1.

Biswas, M., Jibon, A. H., Kabir, M., Mohima, K., Sinthy, R., Islam, M. S. and Siddique, M. (2022). State-of-the-art violence detection techniques: A review, *Asian Journal of Research in Computer Science* **13**(1): 29–42.

Chandane, S., Nadar, A. T., Lokhande, M., Kanthakumar, D. and Shaikh, R. (2024). Violence detection using deep learning, *2024 International Conference on Innovations and Challenges in Emerging Technologies, ICICET 2024*.

Fatima Kiani, G. E. and Kayani, T. (2022). Real-time violence detection using deep learning techniques, *3rd International Conference on Innovations in Computer Science and Software Engineering, ICONICS 2022*.

Gao, Y., Liu, H., Sun, X., Wang, C. and Liu, Y. (2016). Violence detection using oriented violent flows, *Image and Vision Computing* **48-49**: 37–41.

Girsang, A. and Stanley (2023). Cryptocurrency price prediction based social network sentiment analysis using lstm-gru and finbert, *IEEE Access* **PP**: 1–1.

GitHub - seymanurakti/fight-detection-surv-dataset: New generated dataset for fight detection in surveillance cameras. — github.com (n.d.). `https://github.com/seymanurakti/fight-detection-surv-dataset`. [Accessed 11-12-2024].

Kaur, G. and Singh, S. (2022). Violence detection in videos using deep learning: A survey, *Lecture Notes in Networks and Systems*, Vol. 392, pp. 165–173.

Khan, S. U., Haq, I. U., Rho, S., Baik, S. W. and Lee, M. Y. (2019). Cover the violence: A novel deep-learning-based approach towards violence-detection in movies, *Applied Sciences (Switzerland)* **9**(22): 4963.

Mumtaz, N., Ejaz, N., Habib, S., Mohsin, S. M., Tiwari, P., Band, S. S. and Kumar, N. (2023). An overview of violence detection techniques: current challenges and future directions, *Artificial Intelligence Review* **56**(5): 4641–4666.

Negre, P., Alonso, R. S., González-Briones, A., Prieto, J. and Rodríguez-González, S. (2024). Literature review of deep-learning-based detection of violence in video, *Sensors* **24**(12): 4016.

Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A. and Khassanova, M. (2022). State-of-the-art violence detection techniques in video surveillance security systems: A systematic review, *PeerJ Computer Science* **8**: e920.

Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., Ilyas, M. and Mahmood, A. (2019). A review on state-of-the-art violence detection techniques, *IEEE Access* **7**: 107560–107575.

Sernani, P., Falcionelli, N., Tomassini, S., Contardo, P. and Dragoni, A. F. (2021). Deep learning for automatic violence detection: Tests on the airtlab dataset, *IEEE Access* **9**: 160580–160595.

Singh, K., Yamini Preethi, K., Vineeth Sai, K. and Modi, C. N. (2018). Designing an efficient framework for violence detection in sensitive areas using computer vision and machine learning techniques, *2018 10th International Conference on Advanced Computing, ICoAC 2018*, pp. 74–79.

Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., Mostafa, Y. M., Chawky, B. S. and Khattab, D. (2019). Violence recognition from videos using deep learning techniques, *2019 IEEE 9th International Conference on Intelligent Computing and Information Systems, ICICIS 2019*, pp. 80–85.

Sumon, S. A., Goni, R., Hashem, N. B., Shahria, T. and Rahman, R. M. (2020). Violence detection by pretrained modules with different deep learning approaches, *Vietnam Journal of Computer Science* **7**(1): 19–40.

Toma, R., Nahid, A. and Hasan, M. N. (2019). Electricity theft detection to reduce non-technical loss using support vector machine in smart grid.

Ullah, F. U. M., Obaidat, M. S., Ullah, A., Muhammad, K., Hijji, M. and Baik, S. W. (2023). A comprehensive review on vision-based violence detection in surveillance videos, *ACM Computing Surveys* **55**(10).

Wang, P., Wang, P. and Fan, E. (2021). Violence detection and face recognition based on deep learning, *Pattern Recognition Letters* **142**: 20–24.

Yao, H. and Hu, X. (2023). A survey of video violence detection, *Cyber-Physical Systems* **9**(1): 1–24.

Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J. and He, X. (2016). A new method for violence detection in surveillance scenes, *Multimedia Tools and Applications* **75**(12): 7327–7349.