**Violence Detection in Surveillance Video using 3D Convolution network**

Jyoti Kukade, Prashant Panse

[Jyotikukade21@gmail.com](mailto:Jyotikukade21@gmail.com), [Prashant.panse@medicaps.ac.in](mailto:Prashant.panse@medicaps.ac.in)

Medi-Caps University, Indore

**Abstract:**

As the years passed by, computers became more powerful and automation became the need of generation. Humans tried to automate their work and replace themselves with machines. This effort of transition from manual to automatic gave rise to various research fields, computer vision is one such field. Violence detection is an essential task for ensuring public safety in various domains, such as video surveillance, social media monitoring, and crowd management. Detecting violent events accurately and efficiently requires the use of advanced technologies, such as deep learning, which can analyze complex patterns in large datasets. This documentation presents a comprehensive review of the state-of-the-art methods for violence detection using deep learning, including their architectures, training strategies, and evaluation metrics. The documentation starts by discussing the main challenges of violence detection, such as the diversity of violent events, the presence of contextual cues, and the class imbalance problem. Then, it presents the main deep learning architectures used for this task, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their combinations. The documentation describes the design choices of these architectures, such as the number of layers, the activation functions. Moreover, it discusses the training strategies used to optimize these architectures, such as transfer learning, The paper also presents the main evaluation metrics used for violence detection, such as accuracy, precision, recall, and F1 score. In conclusion, violence detection using deep learning is a challenging and essential task that requires the integration of advanced techniques in computer vision and machine learning. The paper provides a comprehensive review of the state-of-the-art methods and datasets for this task and identifies the open challenges and directions for future research.

***Keywords:*** *Computer Vision, Deep Learning, Violence Detection, LSTM, ANN, Action Recognition*

**Introduction:**

Violence is a subjective action, meaning that there is no single set of descriptions for a violent act. In 1996, the World Health Organization (WHO) global consultation on violence and health defined violence as ‘‘the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either result in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation’’. It encompasses various forms, including physical aggression, verbal abuse, and emotional distress, and can lead to both short-term and long-lasting consequences. Consequently, there is a critical need for an automated and intelligent system that can detect violence in real-time and inform emergency services at the appropriate time. The consequences of violence are far-reaching and can affect individuals' physical, mental, and social well-being. For instance, victims of violence may experience long-term health effects, including post-traumatic stress disorder, depression, and anxiety, which can impact their quality of life and ability to function in daily situations. Moreover, violence can have significant economic implications, as it can lead to increased healthcare costs, lost productivity, and criminal justice expenses.

Several surveys and studies have highlighted the harmful impact of violence, especially in real-life scenarios. According to the World Health Organization (WHO), violence leads to significant physical, mental, and social consequences. For instance, a survey conducted by the Centers for Disease Control and Prevention (CDC) in the United States revealed that in 2019, there were approximately 19.2 million reported cases of violence-related injuries. The The National Crime Victimization Survey (NCVS) is a survey conducted by the Bureau of Justice Statistics (BJS) that provides information on criminal victimization in the United States shown in Figure 1, it provides valuable insights into the prevalence and impact of violence, which can be used to inform and improve violence detection models.

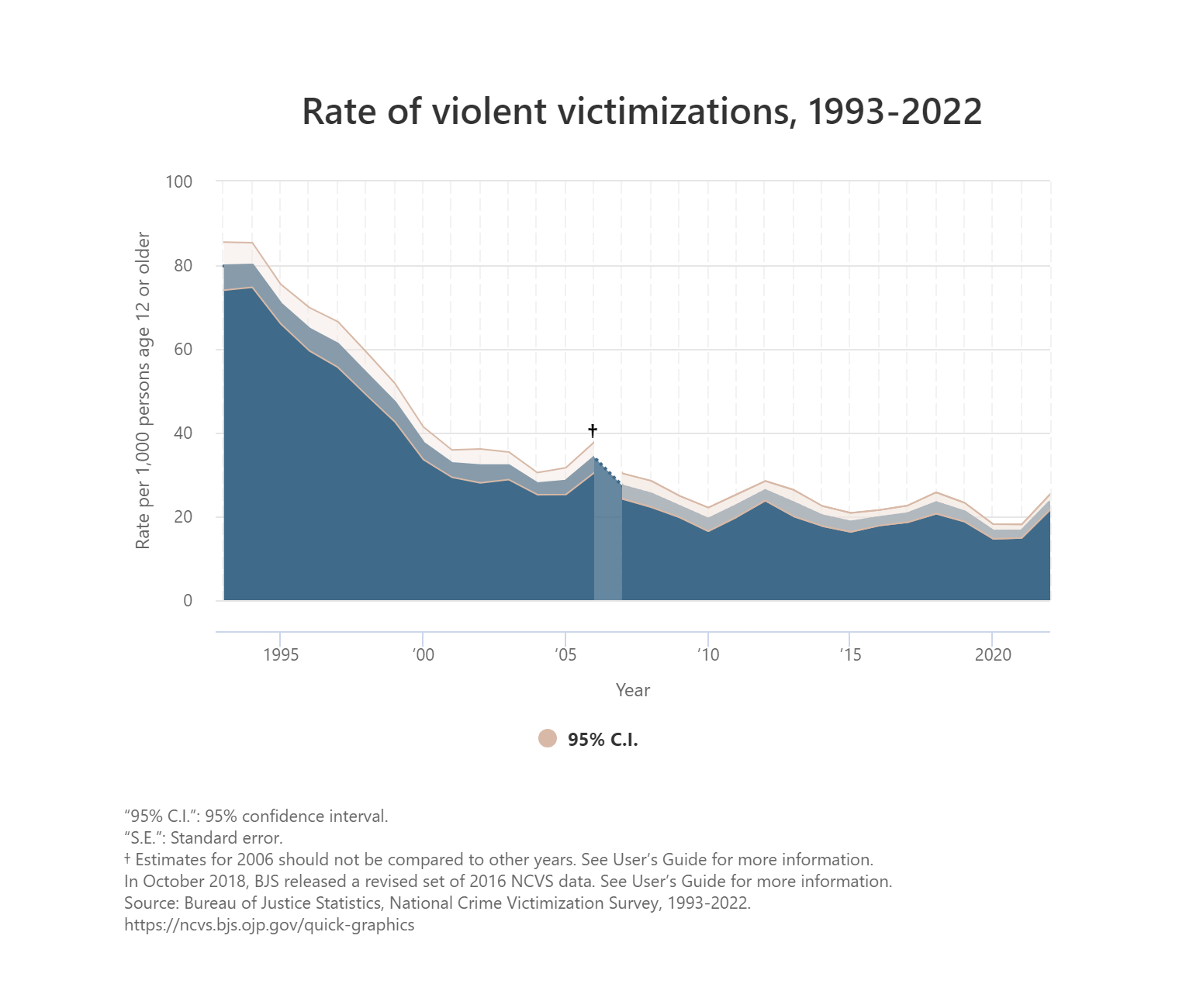


Figure 1: Rate of Violent Victimizations, 1993-2022

Violence Detection plays an important role and is currently getting the attention of researchers because the problem of human action recognition at a distance has become tractable by using computer vision techniques. The recent development in the field of human action recognition has led to a renewed interest in the detection of specific action, one of them is violence. Automatic violence detection is basically involved to increase the safety of humans by monitoring their behavior, in addition to prevent social, economic and environmental damages. Violence detection has become an increasingly important issue in ensuring public safety, particularly in domains such as security, law enforcement, and social media monitoring. Traditionally, detecting violent events has relied on manual annotation and rule-based systems. However, these methods have limited efficacy in handling the complexity and diversity of violent events. Deep learning, on the other hand, has proven to be a promising approach to violence detection. By utilizing large datasets and advanced algorithms, deep learning can analyze complex patterns and learn to detect violent events accurately and efficiently.

Violence detection techniques have been developed to analyze big video data for anomalous actions, ensuring the security of valuable assets and property. However, the challenge of accurately and efficiently detecting violence in real-time surveillance videos remains a pressing issue. This project aims to contribute to the ongoing efforts to develop effective violence detection models in surveillance videos using 3D Convolutional Networks. By leveraging advanced machine learning techniques, it seeks to enhance the accuracy and efficiency of violence detection, ultimately helping to reduce the harm caused by violence in real-life situations.

**Title: Violence detection**

**Related work:**

Violence Detection in Video Using Computer Vision Techniques by E. Bermejo , O. Deniz, G. Bueno, and R. Sukthankar [2] uses two state-of-the-art video descriptors, STIP and MoSIFT, and the bag-of-words framework to represent and classify video sequences as fight or non-fight. It also introduces a new dataset of 1000 hockey video clips containing fights and non-fights. They achieved near 90% accuracy on fight detection on both the hockey dataset and a smaller dataset of 200 action movie clips. It also shows that MoSIFT outperforms STIP on the action movie dataset, indicating its better generalization capacity. The writers come to conclusion that use of local spatio-temporal descriptors such as STIP and MoSIFT, are computationally expensive and require high-resolution videos also more diverse testing and consideration of real-world scenarios would enhance paper’s applicability.

Harnessing high-level concepts, visual, and auditory features for violence detection in videos by Bruno M. Peixoto, Bahram Lavi , Zanoni Dias, Anderson Rocha[3]. The authors evaluate their framework on two public datasets: Hockey Fight and Movies. They compare their results with several baselines and ablation studies. They also conduct a user study to assess the perception and agreement of human annotators on the violence level of videos. They trained different dCNNs (static and motion-based), each of which responsible for detecting a single aspect of violence — focusing on the Mediaeval 2013 VSD task and dataset, which has annotations for several violence concepts. Ultimately, these individual concept classifiers are combined using a tailored network to classify the more general concept of violence, using features from all the networks.

Efficient Violence Detection in Surveillance by Romas Vijeikis, Vidas Raudonis, and Gintaras Dervinis[4],the authors propose a novel and efficient architecture for violence detection from video surveillance cameras, which consists of a spatial feature extractor using a U-Net-like network with MobileNet V2 as an encoder, a temporal feature extractor using LSTM, and a classifier using dense layers. The proposed model is computationally light and still achieves good results—experiments showed that an average accuracy is 0.82 ± 2% and average precision is 0.81 ± 3% using a complex real-world security camera footage dataset based on RWF-2000.

Human skeletons and change detection for efficient violence detection in surveillance videos by Garcia-Cobo, Juan C. SanMiguel [5] proposes a new method for violence detection in surveillance videos that uses human skeletons and change detection. The model's performance is evaluated on three datasets: RWF-2000, Movies, and Crowd. The proposed model demonstrates competitive performance, with an accuracy of 90.25% in the RWF-2000 dataset, 88.00% in the Spatio-Temporal Modeling (Kang et al., 2021) (retrained) dataset, and 62.00% in the Person detector + CNN (Choqueluque-Roman and Camara-Chavez, 2022) dataset. The authors also evaluate their model using a five-fold cross-validation technique to demonstrate the generalization ability of the proposed architecture. The results show that the proposed model is robust and can be applied to various datasets, making it a promising approach for video surveillance applications.

State-of-the-art violence detection techniques in video surveillance security systems: a systematic review by Batyrkhan Omarov, Sergazi Narynov, Zhandos Zhumanov, Aidana Gumar and Mariyam Khassanova [6] emphasizes the importance of evaluating the performance of violence detection models using various metrics, including precision, recall, F1-score, AUC-ROC, false alarm, and missing alarm. The authors provide a detailed explanation of each metric and its significance in evaluating the performance of violence detection models

**Proposed Model**

**Dataset**:

We sourced dataset titled “Real Life Violence Situations Dataset” from Kaggle.[7] Dataset Contains 1000 Violence and 1000 Non-Violence videos collected from YouTube videos. Violence videos comprises mostly street fight, sports fight situations in several 3 environments and conditions. Non-violence videos in the dataset contains videos of different human actions like sports, eating, walking etc. The length of videos varies from 2 seconds to 7 seconds. The dataset was uploaded on April 27, 2019, and is intended for use in automatic violence detection in videos. The videos in the dataset have varying lengths, and the dimensions are not homogenous, including aspect ratios such as 16:9, 3:4, 1:1, and 9:16. The dataset is useful for training and testing models for violence detection in surveillance videos using deep learning techniques such as 3D Convolutional Neural Networks.

**Feature description**:   
The dataset used for the project on Violence Detection in Surveillance Video using a 3D Convolutional Network contains 2000 videos for two classes, with 1000 videos for each class, making it a balanced dataset. The video length varies from 2 to 7 seconds, and the video dimensions are not homogenous, including aspect ratios such as 16:9, 3:4, 1:1, and 9:16.

Several studies and datasets related to violence detection in surveillance videos are available. For instance, the XD-Violence dataset contains 4754 untrimmed films with audio signals and poor labels, while the AICS-Violence dataset includes 7576 high-resolution video clips of violent and non-violent activities. Additionally, a dataset composed of 350 clips labeled as violent and non-violent is also available.

Various techniques and models have been proposed for violence detection, including the use of 3D Convolutional Neural Networks (CNNs) for feature extraction and analysis. These models leverage deep learning approaches for real-time violence detection and have shown promising results in detecting violent activities in surveillance videos. The use of 3D CNNs for feature extraction and analysis has been a common approach in violence detection. These models are designed to capture different relations among video snippets and integrate features, enabling efficient classification across classes.

**Dataset Preparation:**

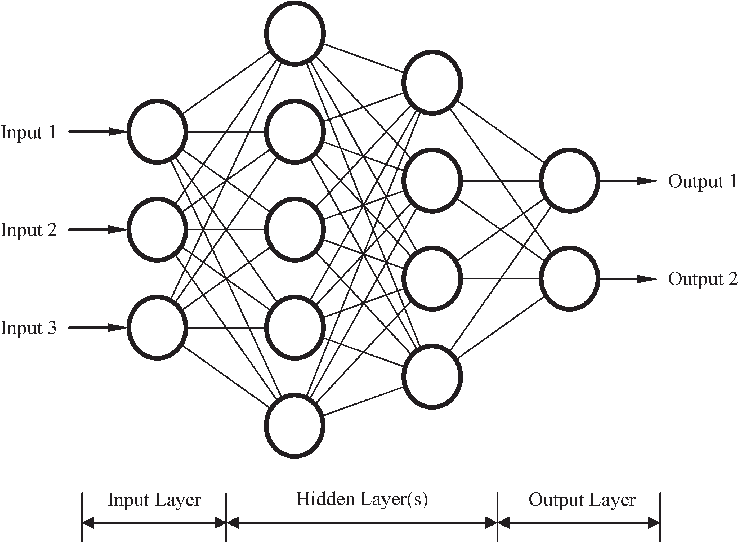
The dataset for violence detection in surveillance video using 3D convolutional networks was meticulously prepared, adhering to a 70:20:10 ratio for training, validation, and testing, respectively. Each video underwent the capture of 12 frames, from which features were extracted using the Inception V3 pre-trained model with ImageNet weights. Subsequently, a feature vector with a shape of (length, 15, 2048) was constructed based on these extracted features, serving as the input for the base models during the training phase. This rigorous dataset preparation process ensured that the 3D convolutional neural network was trained and tested on a well-structured and diverse dataset, ultimately contributing to the accurate detection of violence in surveillance videos.

**Models:**

Different architecture models were used as follows

**Multi-Layer Perceptron**

Multilayer Perceptron (MLP) forms a crucial component. The MLP consists of an input layer and an output layer, both fully connected. However, unlike the perceptron, MLPs may contain multiple hidden layers between the input and output layers.



*Figure 2 MLP Neural Network*

The algorithm for the MLP involves the following steps:

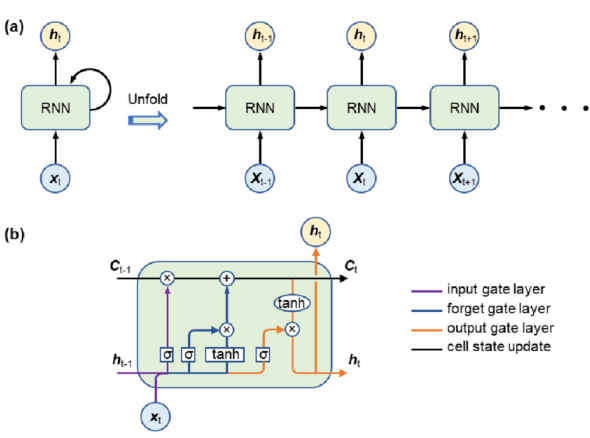
1. The inputs are forwarded through the MLP by computing the dot product of the input with the weights between the input layer and the hidden layer (WH). This yields a value at the hidden layer, which is then passed through an activation function.
2. MLPs utilize various activation functions such as rectified linear units (ReLU), sigmoid function, and tanh. The calculated output at the hidden layer is pushed through one of these activation functions.
3. The output from the activation function is then passed to the next layer in the MLP by computing the dot product with the corresponding weights.
4. Steps two and three are repeated until the output layer is reached.
5. At the output layer, the calculations are used for either a backpropagation algorithm during training or for making a decision based on the output during testing.

MLPs serve as the foundation for all neural networks and have significantly enhanced the computational power for classification and regression problems. They have enabled computers to learn rich and complex models, overcoming limitations such as the XOR cases. In our project, the MLP plays a vital role in the classification of violent and non-violent surveillance videos, contributing to the development of an effective violence detection system.

**Recurrent Neural Network using LSTM layers:**

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are important components. RNNs enable the handling of sequential data by giving the model an internal memory, allowing the input and outputs to loop. The previous input is also considered when making a decision, in addition to the current input and output. LSTMs are extensions of RNNs that have been modified to interpret past data with superior methods.

They consist of three different components or gates: the input gate, output gate, and forget gate. The input gate decides which values are important and should be let through the model, while the output gate decides which values to push to the next time step. The forget gate drops information that the model deems unnecessary to make a decision about the nature of the input values. LSTM neural networks are made up of both special LSTM layers that can interpret sequential data and densely connected layers. Once the data moves through the LSTM layers, it proceeds into the densely connected layers. In our project, LSTMs play a crucial role in interpreting the sequential data of surveillance videos, contributing to the development of an effective violence detection system.



*Figure 3 RNN working and LSTM Cell Architecture*

ConvLSTM

3D Convolution Network

It is to be noted due to computation resource restrictions ConvLSTM and 3D Convolution network were trained on very small part of data and they provided <50% accuracy. Hence, they were not used any further

**Methodology**

The following Fig 4 and Fig 5 visualize the design architecture of the Multi-Layer

Perceptron and LSTM base model.

The pre-trained model works as an feature extractor and takes input vector of shape

(None, MAX\_SEQUENCE\_LEN, IMG\_SIZE, IMG\_SIZE, color\_channel)

Where,

* MAX\_SEQUENCE\_LEN is max number of frames to be captured from the video i.e. 15
* IMG\_SIZE is dimension of frame i.e. 224 in our model
* color\_channel is number of color channel. 3 represents RGB/BGR/GRB channel and1 represents grayscale image.

After each frame is captured and feature are extracted the main input vector is formed for base

Model which has a shape of (Num\_of\_videos, MAX\_SEQUENCE\_LEN, NUM\_FEATURE)

Where,

* Num\_of\_videos is length of dataset.
* MAX\_SEQUENCE\_LEN is max number of frames to be captured from the video i.e. 15
* Since we used Inception V3 model for feature extraction, NUM\_FEATURE is 2048
* Since we used pre-trained model for feature extraction, we saved a lot of computation resources

And time because instead of creating input vector of shape (1796, 15, 224, 224, 3) we passed input

Vector having shape (1796, 15, 2048) which directly reduced trainable parameters and hence the

RAM usage was drastically decreased

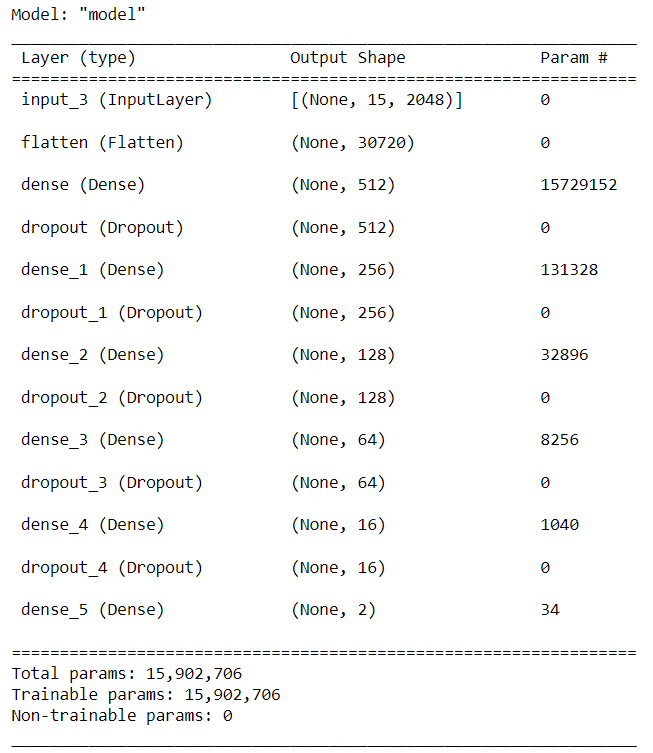


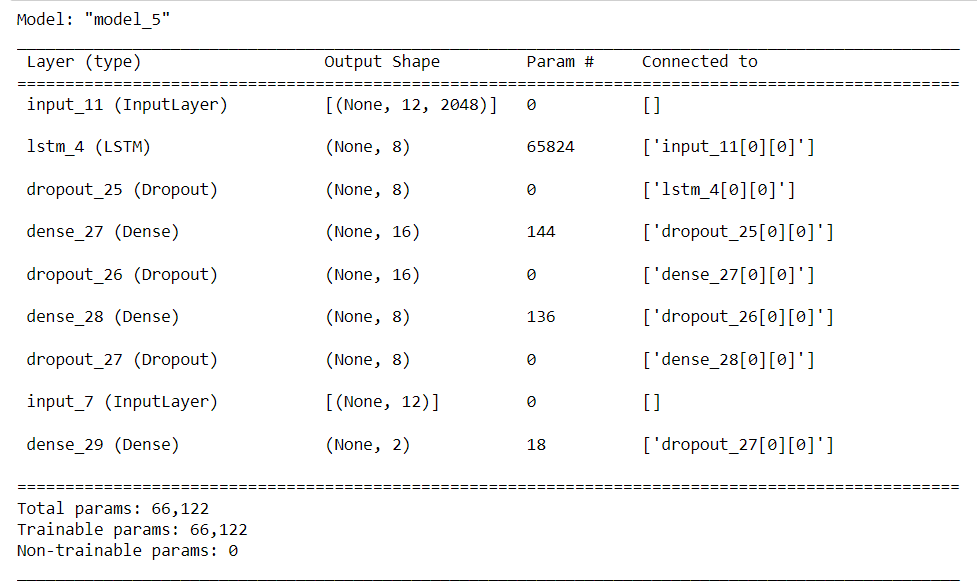
Figure 4 Multi-Layer Perceptron model architecture

We have created an MLP model that takes a 15x2048 matrix of features as input. This matrix can be obtained from a 3D convolutional network or handcrafted features. The MLP model consists of a series of fully connected layers, with dropout layers in between to prevent overfitting. The final layer of the MLP model is a 2-neuron layer that outputs a probability distribution over the two classes: violent and non-violent.

The architecture of our MLP model consists of an input layer, a flatten layer, five dense layers with ReLU activation functions, and five dropout layers. The output layer has two neurons that output a probability distribution over the two classes. To train our MLP model, we need to provide it with a set of labeled training data that contains examples of both violent and non-violent video clips. Once the model is trained, we can use it to classify new video clips as either violent or non-violent.

To classify a new video clip, we first extract the features from the clip using a 3D convolutional network. We then pass these features through the MLP model, which outputs a probability distribution over the two classes. We use this probability distribution to classify the video clip as either violent or non-violent.

MLP models are a simple but effective type of neural network that can be used for violence detection in surveillance videos. Our MLP model is a good starting point for developing a violence detection system using 3D convolutional networks.



*Figure 5 LSTM model architecture*

We have created an MLP model that takes a 15x2048 matrix of features as input. This matrix can be obtained from a 3D convolutional network or handcrafted features. The MLP model consists of a series of fully connected layers, with dropout layers in between to prevent overfitting. The final layer of the MLP model is a 2-neuron layer that outputs a probability distribution over the two classes: violent and non-violent.

The architecture of our MLP model consists of an input layer, a flatten layer, five dense layers with ReLU activation functions, and five dropout layers. The output layer has two neurons that output a probability distribution over the two classes. To train our MLP model, we need to provide it with a set of labeled training data that contains examples of both violent and non-violent video clips. Once the model is trained, we can use it to classify new video clips as either violent or non-violent.

To classify a new video clip, we first extract the features from the clip using a 3D convolutional network. We then pass these features through the MLP model, which outputs a probability distribution over the two classes. We use this probability distribution to classify the video clip as either violent or non-violent.

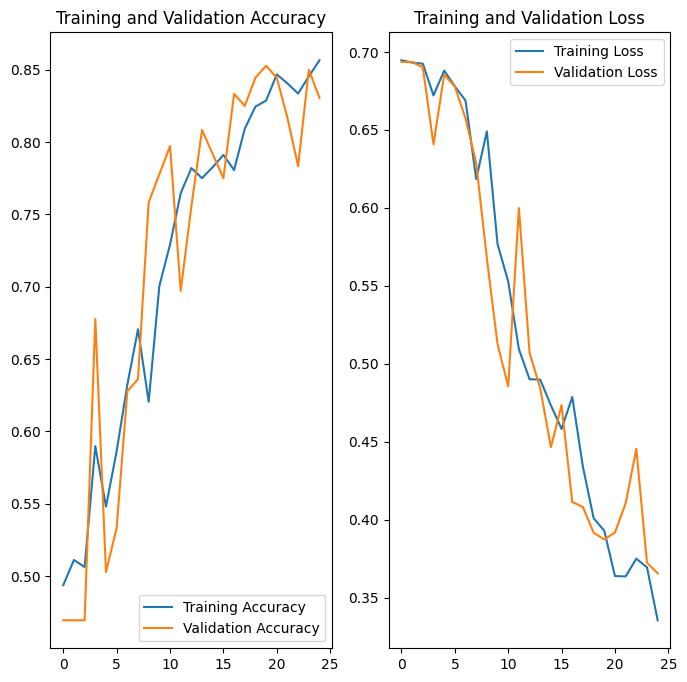
MLP models are a simple but effective type of neural network that can be used for violence detection in surveillance videos. Our MLP model is a good starting point for developing a violence detection system using 3D convolutional networks.

**Training vs validation accuracy and loss:**

**MLP**The accuracy and loss curves in the Fig6 show how well a machine learning model is performing on the training and validation datasets over time. The training accuracy curve shows the percentage of training examples that the model correctly classified, while the validation accuracy curve shows the percentage of validation examples that the model correctly classified. The training loss curve shows the average loss of the model on the training dataset, while the validation loss curve shows the average loss of the model on the validation dataset.

The goal of training a machine learning model is to minimize the loss function, which is a measure of how far the model's predictions are from the true values. As the model learns, the loss function should decrease and the accuracy should increase.

In the Fig6, we can see that the training accuracy and loss curves are both decreasing over time, while the validation accuracy and loss curves are initially decreasing, but then start to increase after epoch 20 where we used early stopping to prevent overfitting.

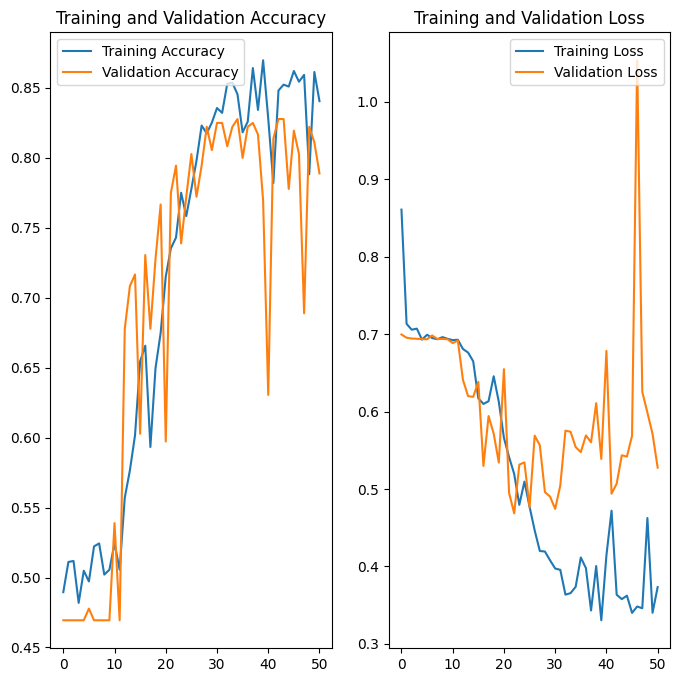
****

*Figure 6 Training VS Validation accuracy and loss plots of MLP*

By analyzing the above plots we can conclude that the our Recurrent Neural Network achieved good fit and provides more than 85+% accuracy on both Training and Validation dataset

**LSTM**

The training and validation accuracy and loss plots of the MLP in the image show how well the model is performing on the training and validation datasets over time. The training accuracy curve shows the percentage of training examples that the model correctly classified, while the validation accuracy curve shows the percentage of validation examples that the model correctly classified. The training loss curve shows the average loss of the model on the training dataset, while the validation loss curve shows the average loss of the model on the validation dataset.

****

*Figure 7 Training VS Validation accuracy and loss plots of LSTM*

By analyzing the above plots we can conclude that the our Multi-Layer Perceptron Neural Network achieved good fit and provides more than 85+% accuracy on both Training and Validation dataset.

**Metrics Report**

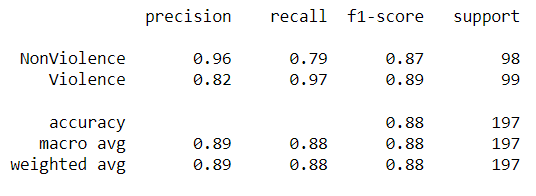
Evaluating the performance of the model using different metrics is integral to every data science project. Here is what you have to keep an eye on: Accuracy is a metric for how much of the predictions the model makes are true. The higher the accuracy is, the better. However, it is not the only important metric when you estimate the performance.

Accuracy is a metric for how much of the predictions the model makes are true. The higher the accuracy is, the better. However, it is not the only important metric when you estimate the performance.

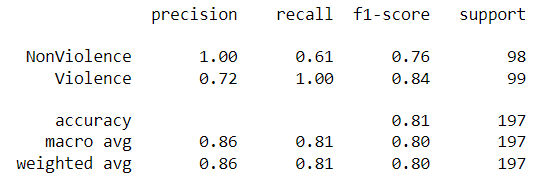
Machine Learning testing scope is determined by the following resultant parameters. Precision should ideally be 1 (high) for a good classifier. Precision becomes 1 only when the numerator and denominator are equal i.e., TP = TP +FP, this also means FP is zero. As FP increases the value of denominator becomes greater than the numerator and precision value decreases (which we don’t want).

Recall should ideally be 1 (high) for a good classifier. Recall becomes 1 only when the numerator and denominator are equal i.e. TP = TP +FN, this also means FN is zero. As FN increases the value of denominator becomes greater than the numerator and recall value decreases (which we don’t want).

So ideally in a good classifier, we want both precision and recall to be one which also means FP and FN are zero. Therefore, we need a metric that takes into account both precision and recall. F1-score is a metric which takes into account both precision and recall and is defined as follows:

******

*Figure 8 Metrics report of MLP model*

**

*Figure 9 Metrics report of LSTM model*

**Confusion matrix of prediction**

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an error matrix. Some features of Confusion matrix are given below:

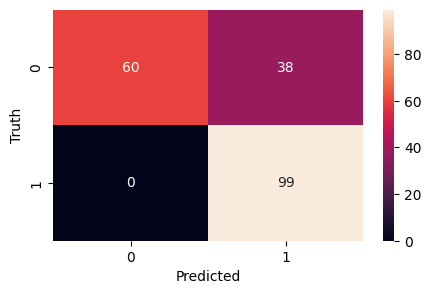
* For the 2 prediction classes of classifiers, the matrix is of 2\*2 table, for 3 classes, it is 3\*3 table, and so on.
* The matrix is divided into two dimensions that are predicted values and actual values along with the total number of predictions.
* Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.
* True Negative: Model has given prediction No, and the real or actual value was also No.
* True Positive: The model has predicted yes, and the actual value was also true.
* False Negative: The model has predicted no, but the actual value was yes, it is also called as Type-II error.
* False Positive: The model has predicted Yes, but the actual value was No. It is also called a Type-I error

The following Confusion matrix represents the performance of our sequential model. The dataset consists of 2 classes hence we have 2x2 matrix. Where 0 = Non-Violence and 1= Violence

*Chart

Description automatically generated*

*Figure 10 Confusion matrix of MLP model*

**

*Figure 11 Confusion matrix of LSTM model*

**Conclusion**

The model trained in this project was trained, tested and validated on dataset consisting of 1996 videos.

The MLP model achieved 87% accuracy and LSTM model achieved 81% accuracy on test dataset which was comprised 198 videos with equal number of videos from each class. The MLP model achieved 0.89 F1 score and LSTM model achieved .84 F1 score.

Multi-Layer Perceptron model performed better than LSTM model.

**References**

[1]<https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5>

[2] <https://www.cs.cmu.edu/~rahuls/pub/caip2011-rahuls.pdf>

[3] https://www.sciencedirect.com/science/article/abs/pii/S1047320321001073

[4] https://www.researchgate.net/publication/359215233\_Efficient\_Violence\_Detection\_in\_Surveillance

[5] https://www.sciencedirect.com/science/article/pii/S1077314223001194

[6] <https://peerj.com/articles/cs-920/>

[7] <https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset>