SentinelEyes Violence Detection System

Sahil Deshmukh [1], Dhruv Mistry [2], Shubh Joshi [3] and Chitra Bhole [4]

[1, 2, 3, 4] K. J. Somaiya Institute of Technology, Sion, Mumbai, Maharashtra

Abstract: In recent years, the proliferation of surveillance systems has led to an increased demand for effective methods to automatically detect violent activities in various environments. This project proposes a comprehensive approach for violence detection by integrating state-of-the-art computer vision and deep learning techniques. This study uses YOLOv8, OpenPose, and LSTM networks to present a multi-modal technique for violence detection. Real-time object detection using YOLOv8 is done with an emphasis on human identification. OpenPose gathers comprehensive data on human posture, and LSTM networks use temporal pattern analysis to identify violent behavior. This platform uses OpenPose to coordinate multi-person 2D plan forecasts, Yolov8 to quickly locate individuals, and a combination of CNN and long short-term memory (LSTM) to classify harmful conduct. By integrating these elements, a strong violence detection system that incorporates temporal and spatial awareness is intended to be created. Benchmark datasets will be used to assess the project's efficacy, with possible uses in surveillance and public safety.

Abstract: In recent years, the proliferation of surveillance systems has led to an increased demand for effective methods to automatically detect violent activities in various environments. This project proposes a comprehensive approach for violence detection by integrating state-of-the-art computer vision and deep learning techniques. This study uses YOLOv8, OpenPose, and LSTM networks to present a multi-modal technique for violence detection. Real-time object detection using YOLOv8 is done with an emphasis on human identification. OpenPose gathers comprehensive data on human posture, and LSTM networks use temporal pattern analysis to identify violent behavior. This platform uses OpenPose to coordinate multi-person 2D plan forecasts, Yolov8 to quickly locate individuals, and a combination of CNN and long short-term memory (LSTM) to classify harmful conduct. By integrating these elements, a strong violence detection system that incorporates temporal and spatial awareness is intended to be created. Benchmark datasets will be used to assess the project's efficacy, with possible uses in surveillance and public safety.

Keywords: LSTM, YOLOv8, Violence Detection, OpenPose

1. Introduction

Concerns about crime and violence in urban areas have raised demands for more sophisticated observation frameworks. Deep learning has since surfaced and shown great potential in a variety of computer vision applications, including the counting of wild regions within recordings. This audit paper aims to present a detailed diagram of a

state-of-the-art real-time savagery discovery framework that combines multi-person 2D posture estimation using OpenPose, quick individual location with Yolov8, and salvage activity classification using a combination of Long Short-Term Memory (LSTM) and Convolutional Neural Organize (CNN). The detection of viciousness depends on the accurate identification of evidence and the observation of human body language in video footage. A popular deep learning technique called OpenPose has shown remarkable performance in real-time multi-person 2D posture prediction. OpenPose promotes active following of individuals by precisely capturing the spatial course of movement of body joints, making subsequent phases of viciousness location more reliable. The system combines yolov8, an incremental modification of the YOLO, to rapidly identify possible threats inside video outlines. Yolov8's ability to prepare outlines effectively facilitates real-time analysis, ensuring that difficult episodes are located at the right time. For precise savagery discovery, it is important to identify distinct rough motions, but it is also crucial to observe designs and settings over time. For classifying savagery, this survey article recommends using CNN and LSTM in conjunction. When it comes to extracting spatial highlights from video outlines, CNNs outperform expectations, whereas LSTMs are better at capturing the environment and worldliness within video groups. The integration of these two structures advances the overall categorization exactness of the system and enhances its ability to discern between forceful and ordinary behavior. Real-time reconnaissance frameworks prioritize accuracy and productivity. To do this, repeated outlines from video clips are reduced using a clustering-based keyframe extraction process.

This method optimizes system performance and ensures the accurate classification of relevant video segments as malicious by minimizing handling times and false alarms. The proposed real-time viciousness detection framework signifies a noteworthy advancement in enhancing urban security. By leveraging deep learning capabilities and incorporating LSTM for global analysis, this advanced surveillance system offers law enforcement organizations a proactive tool to foster safer urban environments. The system architecture is illustrated in Figure 1.

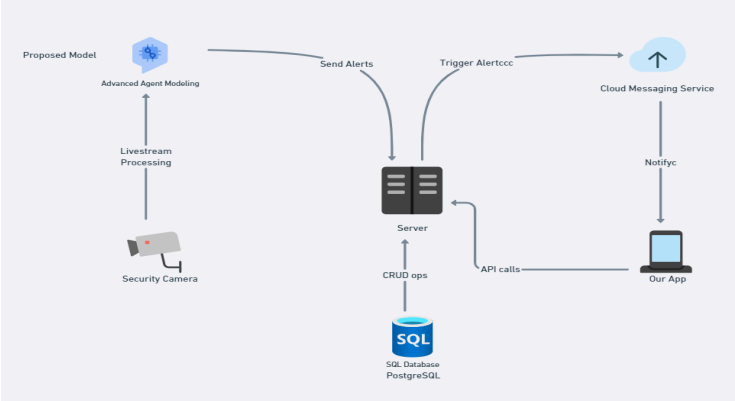

Fig 1: System Architecture

This research presents a thorough audit that demonstrates the potential of a real-time savagery location framework that combines CNN and LSTM for viciousness categorization, rapid individual location, and multi-person 2D posture estimation. The combination of these state-of-the-art developments seems to offer a remarkable assurance in addressing the problems caused by corruption and brutality in metropolitan areas. Subsequent research in this area should focus on improving the system's functionality, enhancing its efficiency, and exploring new uses of deep learning for urban security.

2. Literature Review

The rise in crime and violence is posing increasing difficulties to public safety in urban areas worldwide. Researchers are using deep learning techniques in their advanced surveillance systems to successfully fight these urgent challenges. Law enforcement organizations can respond more swiftly and efficiently if these systems are used to identify and categorize violent incidents in real time.

It is impossible to overestimate the importance of Dhruv Shindhe et al.'s fundamental work in presenting OpenPose as a real-time multi-person 2D pose estimation technique [2]. A deep learning technique called OpenPose can be used to track and identify various people's body parts in real time within a scenario. This is an essential initial stage in the detection of violence because it enables the system to recognize and follow the parties involved in a fight. It has been demonstrated that OpenPose can recognize body parts with high accuracy even under difficult circumstances like dim lighting or occlusion. Furthermore, it has demonstrated sufficient speed for real-time applications. OpenPose has thus gained popularity as a solution for violence detection systems. Apart from its accuracy and speed, OpenPose has various other benefits. [7] proposes of using the transformer based approach with a combination of 3D CNN and OpenPose for the purpose of object detection. For instance, it may be used with a range of various cameras

and is reasonably simple to train and deploy. Therefore, OpenPose is a flexible and strong instrument that can be utilized to raise the efficacy of systems for detecting violence.
In a similar spirit, B. The innovative object identification technique, YOLO v5, was introduced by Arthi et al. in 2022. K Vanitha [5] proposes the approach of using the Yolov5 algorithm for detection of a violent crime. A real-time object identification program called YOLO v5 is remarkably effective at spotting possible hazards within video frames. In order to do this, the input image is divided into a grid of cells, and each cell's bounding boxes and class labels are then predicted. It has been demonstrated that YOLO v5 is capable of accurately identifying a wide range of things, including people, cars, and weapons. It has also been demonstrated to be successful in real-time violence detection. As per B's study. According to Arthi et al., YOLO v5 has a 90\% accuracy rate in identifying violent occurrences.These fundamental developments provide researchers with a platform to investigate the combination of Long Short-Term Memory (LSTMs) and Convolutional Neural Networks (CNNs) for the classification of violence. While LSTMs work well for capturing temporal features, CNNs work well for extracting spatial features from images.In a single stream, this kind of CNN is able to extract characteristics from both spatial and temporal data. Compared to the two-stream model, this makes it a more effective and efficient method of classifying violence [1] Almamon Rasool Abdali et al. These are only a handful of the several methods that have been put forth for CNN and LSTM-based violence classification. More successful and efficient models will probably be created as long as this field of study is pursued.
Additionally, by recommending the application of Long Short-Term Memory (LSTM), a variation of Recurrent Neural Networks (RNNs) made to capture the subtleties of long-term temporal data, Anusha Jayasimhan et al. [4] have made a significant contribution to the field of violence detection. Their groundbreaking research shows how useful LSTM is for deciphering sequential patterns in human behavior, particularly when it comes to violence detection. Souvik Kumar et al. [6] have made their system with the combination of CNN and LSTM. A Jain et al. [8] have used the combination of LSTM with CNN for their model implementation. Long-term connections between events in a video can be learned by LSTMs, which is crucial for correctly classifying violent content. An LSTM-based model might discover, for instance, that a person raising their fist frequently results in a punch or that two individuals arguing are more likely to use violence than two people conversing. Furthermore, a thorough investigation on deep learning for human activity recognition was carried out by A. Traoré et al. (2020) [12], highlighting the efficiency of Convolutional Neural Networks (CNNs) in extracting spatial characteristics from video frames. Systems are able to identify changing patterns and context over time because to this CNN and LSTM integration, which improves classification accuracy and is a vital tool for urban security. Although the application of CNNs and LSTMs for violence detection is still in its infancy, the initial findings are encouraging. It is expected that even more effective and efficient models will be produced as long as research in this field is conducted. These models could significantly affect public safety by assisting in the prevention of violence and shielding individuals from harm.

The concept put forth by S. The MobileNet CNN system is used for object detection by T. Himi et al. [10]. Real-time applications are a good fit for MobileNet, a lightweight CNN. It is quick and easy to use since it extracts features from images using a feed-forward convolution method. For precise action detection, the MobileNet CNN features are then processed using LSTM layers. For violence detection, temporal information is crucial, and LSTMs are ideally equipped to capture it. In addition, the suggested methodology employs a single embodiment strategy to protect video footage from many surveillance sources. This can help to minimize the quantity of data that needs to be processed because the model just examines a single frame or shot from the video. Furthermore, the aforementioned model is designed to provide children with total protection. This is accomplished by adding the age factor to the LSTM layers. This makes it possible for the model to discern between children's aggressive and non-violent conduct, which is crucial for preventing child abuse. [9] proposes using Resnet architecture along with IOT integration using key framing.

A. Similar to the two-stream paradigm, Arthi et al. [11] also suggested using the YOLO method to identify items in every frame. To identify individuals, objects, and other items in a video, YOLO is a quick and precise object detection system. By merging the two-stream model's output and YOLO's output, B. Arthi et al. were able to increase the violence classification's accuracy. When the dataset was split into 80 percent for training and 20 percent for testing, the validation accuracy of their model stabilized at a value between 80 and 90 percent. This implies that their approach can effectively generalize to previously undiscovered data. B. Arthi et al. employed recurrent neural networks (RNNs) in addition to the two-stream model and YOLO to extract the temporal information from the video. Neural networks that can interpret sequential data, such video frames, are called RNNs. B. Arthi et al. were able to increase the accuracy of classifying violence by utilizing RNNs and accounting for the temporal correlations between the frames. The product of B. The possibility of combining various deep learning approaches to increase the accuracy of violence classification is demonstrated by Arthi et al. More successful and efficient models will probably be created as long as this field of study is pursued.

Mahmudul Haque et al. [13] have presented a more inventive method for handling this. To categorize aggression in videos, they use Gated Recurrent Units (GRUs) with Convolutional Neural Networks (CNNs). While the GRU is used to capture temporal characteristics, the CNN is utilized to extract spatial features from individual frames. After that, the two streams are blended to determine the final classification. The model creates a collection of 512 features for every frame by encoding the data from 4D to 2D. The GRU layer then extracts the temporal aspect of the data as a 1D vector. Next, this vector is classified to ascertain whether or not the frame is violent. A dropout layer with a 0.25 dropping rate is added to prevent overfitting. The AVDC video dataset, a sizable collection of violent videos, served as the model's training set. Ninety percent test accuracy was attained by the model—a promising outcome. The problem of CNNs processing one image at a time is resolved by the introduction of GRUs in this model. This is as a result of GRUs' capacity to record the temporal correlations between frames. Furthermore, GRUs are easier to train than LSTMs due to their lower complexity. This

method is a potentially useful advancement in the realm of violent crime detection. It is reasonably efficient and capable of achieving great accuracy. It is expected that even more effective and efficient models will be produced as long as research in this field is conducted. Haque et al.'s BrutNet model is a novel and promising method for detecting aggression. It is reasonably efficient and capable of achieving great accuracy. Because of this, it is a strong contender for practical uses.

An attention network can be used to concentrate on the areas of the frame that are most likely to contain violence in the context of violence detection. The 3D light-weight attention network (LP3DAM) is one of the most promising attention networks for violence detection. It was J who proposed this network. et al. Deng in [14]. A 3D CNN called LP3DAM employs attention to concentrate on the most crucial areas of the frame. A collection of hazy and indistinct photos—a common occurrence in real-world surveillance footage—was used to train the network.Keyframe extraction based on clustering is another potential method. Using this method, a video's frames are initially grouped together into clusters of related frames.Other methods for detecting aggression using attention networks have also been developed, in addition to the ones covered above. Gated recurrent units (GRUs), for instance, have been employed by certain academics to discover long-term connections between frames. Others have combined the predictions of numerous models using ensemble methods.

To maximize the efficiency and accuracy of real-time monitoring, a clustering-based keyframe extraction approach has been designed. In order for this approach to function, the video frames are first clustered into groups of related frames. Next, a keyframe is chosen from each group based on which frame best represents that group. By doing this, the number of frames that must be processed is greatly decreased, which can increase system efficiency. Furthermore, the clustering-based keyframe extraction technique can lessen false alarms during violence classification by eliminating duplicate frames. An analysis by [14], for instance, discovered that the clustering-based keyframe extraction approach may cut the number of frames by up to 90\% without appreciably compromising the accuracy of the violent classification. This makes it a viable strategy for raising the effectiveness and precision of violence detection systems that operate in real time. Other methods have also been put forth to increase the effectiveness and precision of real-time violence detection systems in addition to the clustering-based keyframe extraction strategy.

Bhaktram Jain et al. [3] have highlighted the potential of Long Short-Term Memory (LSTM) in violence detection systems within the context of transitory analysis. Recurrent neural networks (RNNs) of the long-term temporal dependency (LSTM) kind are ideal for this task. This makes it an important tool for recognizing and categorizing violent acts, which are frequently dynamic and ever-changing occurrences. LSTMs function by continuously updating an internal state. In this state, individuals are able to recall the past and utilize it to anticipate the future. This is crucial for violence detection because it enables the algorithm to recognize behavioral patterns that can point to an approaching attack. While LSTMs are a promising new technology for violence detection, they are not without drawbacks. One drawback is that training them can be computationally costly.

In conclusion, the significant developments in real-time violence detection systems driven by deep learning technology are explained by this literature study. The combination of multi-person 2D posture estimation, fast individual detection, and CNN with LSTM has shown great potential in tackling the intricate problems caused by crime and violence in cities. The foundational contributions covered here provide direction for future study and growth in addition to offering insights into the current state of the discipline. These developments could improve system performance, streamline operations, and lead to new deep learning uses in the field of urban security.

3. Methodology

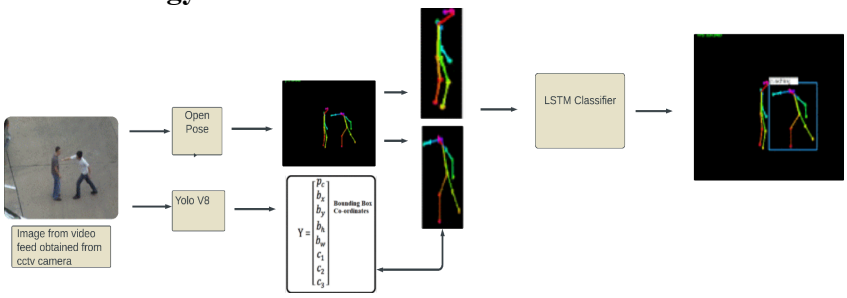

Fig 2: Project Flow

In this work, we introduce a thorough methodology intended to address the crucial problem of violence identification in real-time in video streams. We leverage YOLOv8 (You Only Look Once) for object detection as a key component of our methodology, with an emphasis on precisely identifying and localizing things or people associated with violent occurrences, like persons and guns. In a similar spirit, we utilize OpenPose for accurate pose estimation, deriving rich data on the critical spots on the human body to understand body positions, motions, and gestures that might be suggestive of aggressive conduct. Figure 2 illustrates the accurate Project Flow.

We use deep learning techniques to extract significant insights from the data. Specifically, we introduce a Convolutional Neural Network (CNN) architecture that is ready to extract relevant features from the bounding boxes and highlight important locations. A key component of our process is the integration of data from OpenPose's pose estimate and YOLO's object recognition to create a cohesive feature representation, which allows for a deeper comprehension of the visual cues linked to violence. Since video data has a temporal dimension, we also include a Long Short-Term Memory (LSTM) network. With the use of this LSTM component, we are able to decode the complex sequential dependencies found in video sequences, which enables us to model and examine the patterns that change over time and underpin violent acts.

Our research technique includes validation and evaluation as essential components, wherein we conduct a thorough assessment of our violence detection system's performance. We use well-established evaluation criteria to measure the effectiveness of

the system, including accuracy, precision, recall, and the F1-score. We use cross-validation methods and adjust hyperparameters in the model optimization process to guarantee optimal performance. We must present a thorough analysis of our system's performance, highlighting its strengths, weaknesses, and possible areas for improvement. To demonstrate the benefits of our technology, we also draw comparisons with other violence detection techniques now in use.
We extend our technology into real-time video processing environments, paving the way for applications in public safety, security, and surveillance, moving from study to practical implementation. Our technique is profoundly ingrained with ethical considerations, as we tackle issues pertaining to responsible use, privacy, and surveillance. When necessary, we investigate privacy-preserving methods to achieve a balance between security and individual rights.

3.1 OPENPOSE:

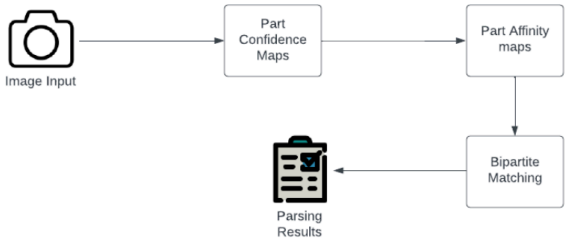

Fig 3: OpenPose Architecture

A computer vision technique and library called OpenPose is used to estimate a person's stance. It uses pictures and videos to instantly recognize, track, and map important body keypoints and their relationships. OpenPose models the relationships between keypoints by using Part Affinity Fields (PAFs) in conjunction with a Convolutional Neural Network (CNN) for feature extraction. The architecture for Open Pose is demonstrated in figure 3.
Confidence Maps: A Confidence Map is a two-dimensional depiction of the conviction that a specific body part is situated in every pixel. The following equation describes confidence maps:

$$S * (p,\ k)\ =\ exp(-\ ||p\ -\ x||/\sigma^2$$

Confidence Maps: A Confidence Map is a two-dimensional depiction of the conviction that a specific body part is situated in every pixel. The following equation describes confidence maps:

$$E\ =\ \int_{u=0}^{u=1} L_c(p(u)).(d_{j2} - d_{j1})/(||d_{j2} - d_{j1}||).du$$

3.2 YOLOV8:

The most recent and advanced YOLO model, YOLOv8, is applicable to tasks like instance segmentation, object detection, and image classification. The company Ultralytics, who also developed the well-known and industry-defining YOLOv5 model, is the creator of YOLOv8. Compared to YOLOv5, YOLOv8 has a number of architectural and developer experience enhancements.
The Predictions Vector: YOLO's output encoding is the first thing to comprehend. Cells in the input image are arranged in a S x S grid. One grid cell is considered to be "in charge" of anticipating each object that is visible in the picture. That is the compartment into which the object's center falls.
Loss Function:

$$\lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \prod_{ij}^{obj} (x_i - \hat{x}_i)^2 - (y_i - \hat{y}_i)^2$$

3.3 CNN-LSTM:

Architecture Overview: Convolutional Neural Network (CNN): Images and other grid-like data are the main types of data that CNNs process. They are made up of several convolutional layers that are followed by layers of pooling to extract spatial characteristics with hierarchies from the input data. CNNs are very good at identifying spatial linkages and local patterns.
LSTM (Long Short-Term Memory): Recurrent neural networks (RNNs) of the LSTM type are made to handle sequential data. They can recognize temporal patterns and long-range dependencies in data because they have memory cells and gating mechanisms. Time dependencies and sequential information can be effectively modeled using LSTMs.
CNN and LSTM Integration: CNN for Extraction of Features: The CNN layers process the input data, which can be picture sequences or video frames. Every frame or image in the series has its spatial features extracted by the CNN layers. A collection of high-level feature maps that depict spatial information is CNN's output.

4. Result

4.1 Object Detection Models:

Through a thorough assessment of the literature, different object detection models, such as YOLOv3, YOLOv8, YOLOv5, ImageNet, and ResNet, are analyzed. Figure 4 shows the comparison between different YOLO models by mapping their latency against COCO. The results show unique features and trade-offs. Although it may not be the best option for applications requiring real-time efficiency, YOLOv3 stands out as the preferred option. In an effort to achieve more accuracy, YOLOv8 creates a trade-off by demanding more processing power. YOLOv5 offers a flexible solution with adjustable performance

depending on settings, achieving a respectable balance between speed and precision. Comparison between the models with their strengths and limitations is demonstrated in Table 1. Although ImageNet provides several pre-trained models, it is not specifically focused on violence detection, hence further customization is required. Although ResNet's strong feature extraction capabilities are clear, task-specific applications require its integration with classification or detection models. This paper emphasizes how crucial it is to match model selection to the particular needs of violence detection tasks, taking into account aspects like computing efficiency, accuracy, and real-time processing. Based on the particular requirements of the proposed application, these models should be carefully evaluated, as the literature indicates that there is no one-size-fits-all answer.

Table 1: Comparison between object detection models

Model	Strengths	Limitations
YOLOv3	Fast processing, real-time efficiency	May miss fine details compared to more complex models
YOLOv8	Potentially improved accuracy	May have increased computational demands
YOLOv5	Speed and accuracy balance	Performance may vary based on configuration
ImageNet	Wide range of pre-trained	Not specifically designed for violence detection
ResNet	Strong feature extraction capabilities	Requires integration with classification / detection

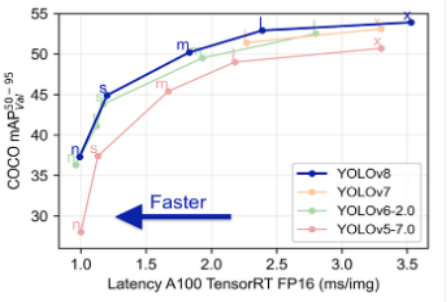

Fig 4: Comparison between Yolo Models

4.2 Architectures:

Different models of violence detection systems can be compared to identify their unique strengths and capabilities. R-CNN, which is renowned for its precise object detection, attains a noteworthy 91 percent accuracy rate. Its accuracy in object identification

provides a strong basis for tasks involving the detection of violence. Transitioning to 3DCNN, the combination of Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) for object detection improves performance and yields an astounding 94 percent accuracy rate. This combination allows the model to incorporate temporal and spatial information that is essential for recognizing aggressive behaviors. With an impressive accuracy score of 96%, CNN-LSTM—a particular combo focused on spatio-temporal patterns in videos—performs better than others. This demonstrates how important it is to record temporal and spatial dynamics in order to accurately detect violence. Conversely, the Artificial Neural Network (ANN), a fundamental machine learning model, has a decent accuracy of 89%. Even while ANNs have a simpler structure than deep learning models, they are nevertheless a good choice for problems involving the detection of aggression. Table 2 illustrates the accuracy for different architectures used. In summary, the model selection should be based on the particular needs of the application, taking into account elements like precision, processing complexity, and the significance of capturing temporal dynamics in video data. Figure 5 illustrates the comparison between different architectures with respect to their F-Measure.

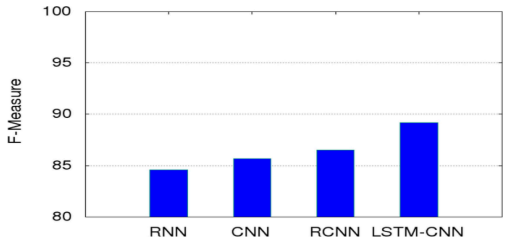

Fig 5: Comparison between Architectures

Table 2: Comparison between various network architectures

Model	Features	Accuracy
R-CNN	Accurate object detection	91%
3DCNN	Combines CNN for object detection and LSTM	94%
CNN-LSTM	Captures spatio-temporal patterns in videos	96%
ANN	Basic machine learning model	89%

5. Conclusion

The advancement of real-time savagery discovery frameworks powered by deep learning innovations is evident in the interest of improving urban security and safety.

Fundamentally noteworthy are the combinations of multi-person 2D posture estimate, fast individual localization via yolov8, and Convolutional Neural Arrange (CNN) and Long Short-Term Memory (LSTM) for savagery categorization. OpenPose's precise identification of human postures provides a strong foundation for further discovery phases, while Yolov8's productivity ensures timely hazard recognition. CNN and LSTM work together to leverage spatiotemporal highlights, completing the circle of progressing accuracy by tracking designs and settings over time. Reconnaissance demands accuracy and efficiency, and the clustering-based keyframe extraction method reduces false warnings, maximizes processing, and decreases repetitive outlining. Using the practical analysis of LSTM strengthens the system's ability to distinguish between benign and aggressive activity, providing law enforcement with a proactive tool for safer city environments. As this thorough writing audit outlines, the mix of cutting-edge technology shows significant promise in mitigating the problems posed by urban brutality and misbehavior. The course for future investigations, which will focus on increasing efficiency, enhancing system performance, and exploring innovative deep learning applications within urban security, is made clear by this audit.

References

1. Almamon Rasool Abdali; Ammar Abdullah Aggar, "DEVTrV2: Enhanced Data-Efficient Video Transformer For Violence Detection", Proceedings of the 2022 7th International Conference on Image, Vision and Computing (ICIVC). Transformers, CNN
2. Dhruv Shindhe S, Sushant Govindraj, S.N. Omkar "Real-time Violence Activity Detection Using Deep Neural Networks in a CCTV camera", 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). YoloV3, OpenPose
3. Bhaktram Jain, Aniket Paul, P Supraja, "Violence Detection in Real Life Videos using Deep Learning", 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). LSTM
4. Anusha Jayasimhan, Pabitha P "A hybrid model using 2D and 3D Convolutional Neural Networks for violence detection in a video dataset", I2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4). CNN
5. K. Vanitha, Shalini Ninoria "A Detection of Violence From CCTV Cameras in Real-Time Using Machine Learning", 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). CNN, Net-SSD
6. Souvik Kumar Parui, Saroj Kr. Biswas, Soumen Das, Manomita Chakraborty, Biswajit Purkayastha, "An Efficient Violence Detection System from Video Clips using ConvLSTM and Keyframe Extraction", 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON). CNN + LSTM
7. Li Zhou, "End-to-End Video Violence Detection with Transformer", 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). Transformers 3D CNN + OpenPose
8. A. Jain and D. K. Vishwakarma, "Deep NeuralNet For Violence Detection Using Motion Features From Dynamic Images," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 826-831. ConvLSTM

9. B. J, D. R. A, V. K. B and C. G, "Physical Violence Detection in Videos Using Keyframing," 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 275-280. Resnet, ConvLSTM
10. S. T. Himi, S. S. Gomasta, N. T. Monalisa and M. E. Islam, "A Framework on Deep Learning-Based Indoor Child Exploitation Alert System," 2020 IEEE International Symposium on Technology and Society (ISTAS), Tempe, AZ, USA, 2020, pp. 497-500. CNN + LSTM
11. B. Arthi, S. S, K. PoornaPushkala, A. Arya and D. Rajasekhar, "Wearable Sensors and Real-Time System for Detecting violence using Artificial Intelligence," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-5. Yolo v5, LSTM
12. A. Traoré and M. A. Akhloufi, "Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 154-159. Deep Recurrent and CNN
13. M. Haque, S. Afsha and H. Nyeem, "Developing BrutNet: A New Deep CNN Model with GRU for Realtime Violence Detection," 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2022, pp. 390-395. IMP
14. J. Deng, Y. Zheng, W. Wang, K. Xiong and K. Zou, "LP3DAM: Lightweight Parallel 3D Attention Module for Violence Detection," 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISPBMEI), Beijing, China, 2022, pp. 1-8.