**Violence Detection System Using MobilenetV2 and Frame Extraction**

***A Smart Cities Security Implementation***

Swarnalatha P, Surya Singha, Srajal Agarwal, Harshit Jain

**Abstract**

The violence detection system has been built on the framework of CNN and LSTM, as to provide a well working solution for bettering public safety by making the detection of violence through CCTVs AI integrated so as to make the processing of highly dynamic environment and data. To do this we use the power of CNNs to extract special features from video provided and LSTM networks are used to analyse temporal patterns that form over time to help the model learn and get better. Such technologies are adept at recognizing the dynamic world conditions and can also support the existing CCTVs AI integrated so as to make the processing of highly dynamic environment and data. To do this we making it ideal for modern smart city applications.

**Introduction**

In todays every changing world and landscape the safety of its people has become a prority and particularly in public spaces where large violent activities can take place a lot of times and these are more often than not unpredictable, this is where traditional survailance systems fail as they lie purely on human input which has a huge downsides.

These downsides include but are not limited to human error and fatigue over longer periods of observation. Such challenges call for there to be a better way to be present for survailence. This is where AI comes in to help us as these pre trained models can go through thousands of frames of data tirelessly and using machine learning and deep learning techniques to predict when and where there is going to be a violent activity taking place. They not only improve the accuracy and efficiency but also reduce the human operators and hence bringing down the overall costs to run these tasks significantly. This capability makes them an invaluable component in the development of smart cities, where technology-driven safety measures are essential for fostering trust and well-being among citizens.

**Literature survey**

**1: Designing an Efficient Model for Violence Detection Using Advanced Deep Learning Techniques**

**- Inference/Achieved:** Inference: This study provides a model that leverages 3D CNNs for undertaking the challenge of violence detection in videos in real time. 3D CNNs provide a much more in depth analysis of the video to get us a better feature extraction as it goes in depth looking into special as well as temporal information. MobileNet has also been integrated which a lightweight and efficient neural network is allowing the 3D CNN to focus on critical details indicative of violence, making it a viable option for real-time implementation.

**- Gaps Identified:** The model faces issues such as bottlenecks due to high memory, changing camera angles, and scalability issues. Relying on traditional inference methods also limits its adaptability to complex real-world situations.

**2: Real-Time Violence Detection Using Deep Learning Techniques**

**- Inference/Achieved:** This research introduced an innovative hybrid model combining AlexNet and SqueezeNet architectures, with ConvLSTM as a temporal analysis layer. The choice of AlexNet provided depth and feature extraction power, while SqueezeNet added computational efficiency, making the model lightweight and suitable for real-time applications. ConvLSTM facilitated the extraction of temporal relationships between video frames, enabling the detection of events as they unfolded over time. The fusion of these models was meticulously designed to take advantage of their complementary strengths, resulting in high accuracy and low latency on benchmark datasets. The system was also tested under varying conditions, such as different lighting and crowd densities, demonstrating its robustness. This approach underscores the importance of hybrid architectures in solving complex tasks like violence detection, where both spatial and temporal cues are critical.

**- Gaps Identified:** The study highlights a reliance on labelled training data, limiting scalability to diverse real-world environments. Challenges in deployment under uncontrolled conditions, such as dynamic public spaces, remain unresolved.

**3: An Effective Approach for Violence Detection Using Deep Learning and Natural Language** **Processing**

**- Inference/Achieved:** This study explored the integration of deep learning with Natural Language Processing (NLP) to enhance multi-modal violence detection. The proposed system combined visual features from deep learning models with textual cues derived from NLP, offering a richer context for scene understanding. For example, it analysed not only the physical actions captured in video frames but also the accompanying audio or textual context, such as speech or environmental sounds, to identify violence. This dual-stream approach proved effective in situations where visual cues alone were insufficient to determine violent activity, such as verbal altercations without visible physical aggression. By fusing these modalities, the system achieved higher accuracy and demonstrated the potential for addressing ambiguities inherent in single-modal systems. The research highlights the growing importance of multi-modal frameworks for complex real-world applications, such as smart surveillance systems.

**- Gaps Identified:** Synchronization of video and textual data streams in real time remains a significant challenge. The study also lacked specific details on how the model could handle latency issues or scale to large, dynamic environments.

**4: Efficient Aggressive Behaviour Detection and Alert System Employing Deep Learning Techniques**

**- Inference/Achieved:** The authors of this study utilized an ensemble approach that combined multiple deep learning models to detect aggressive behaviour by analysing audio and visual inputs simultaneously. This methodology allowed the system to draw from complementary information sources, such as detecting shouting from audio streams and aggressive gestures from video footage. By leveraging ensemble learning, the system achieved improved generalization and performed well across diverse datasets, demonstrating its ability to handle varying types of violent behaviour. The use of multi-modal data further enhanced the system’s robustness, making it more effective in identifying subtle or ambiguous signs of aggression that might be missed when relying on a single data source. This work underscores the importance of ensemble models in achieving higher accuracy and reliability in violence detection tasks, especially in dynamic, real-world environments.

**- Gaps Identified:** The study identified a need for more diverse datasets and improved handling of false positives, particularly in scenarios with ambiguous or borderline aggressive behaviours.

**5: A Review on State-of-the-Art Violence Detection Techniques**

**- Inference/Achieved:** This review offered a detailed examination of recent advancements in violence detection, with a focus on how state-of-the-art deep learning methods are applied to both spatial and temporal analysis. The paper outlined various approaches, such as the use of CNNs for spatial feature extraction and RNNs or LSTMs for temporal analysis, emphasizing their effectiveness in controlled experimental settings. The review also highlighted the growing role of hybrid architectures that combine multiple models to address the challenges of real-time detection. By summarizing the strengths and limitations of current techniques, the paper provided a comprehensive overview of the field, making it a valuable resource for understanding the current landscape of violence detection research.

**- Gaps Identified:** A recurring challenge noted was the lack of datasets that comprehensively represent real-world violence scenarios, limiting the models’ applicability outside controlled environments. Issues such as occlusions, lighting variations, and high false-positive rates remain barriers to practical implementation.

The reviewed studies collectively highlight the advancements in deep learning-based violence detection systems while also exposing critical challenges in their practical implementation. One of the key achievements is the development of sophisticated architectures, such as 3D CNNs and hybrid CNN-LSTM models that capture both spatial and temporal features. These approaches have demonstrated high accuracy and efficiency, particularly in controlled experimental settings. For instance, 3D CNNs effectively analyse video sequences by integrating spatial and temporal information, and the use of lightweight networks like MobileNet enhances computational efficiency for real-time applications. Similarly, hybrid models combining AlexNet and SqueezeNet with ConvLSTM have leveraged complementary strengths for robust and low-latency violence detection. Multi-modal systems further extend these capabilities by integrating visual, audio, and even textual data, offering richer contextual understanding and improved detection accuracy, especially in ambiguous scenarios.

**Challenges in Violence Detection**

Violent incidents in urban environments, including physical altercations, vandalism, and aggressive behaviours, present significant challenges for real-time response. Manually monitoring thousands of CCTV feeds is impractical . This project aims to bridge the gap by developing a deep learning-based violence detection system capable of analyzing video footage for violent activities and sending alerts in real-time.

Violence detection systems face several challenges that impact their performance . Environmental variability, such as changes in lighting conditions, weather, and crowd density, significantly affects detection accuracy, as models struggle to generalize across diverse scenarios. Additionally, real-time processing requires high-speed, low-latency computations, which can be resource-intensive and difficult to achieve. Another problem is the occurrence of false positives and false negatives as they also hinder operational efficiency, leading to delays or overlooked incidents.

Integration with existing smart city infrastructure and security systems poses a challenge, as compatibility is crucial for ensuring scalability and adoption in urban environments. Addressing these challenges is essential for creating a stable and usable violence detection solution.

**Methodology and Architecture**

Our solution incorporates several unique features that enhance its effectiveness and suitability for real-world violence detection. It leverages a hybrid CNN-LSTM architecture, where Convolutional Neural Networks (CNNs) are responsible for extracting spatial features, while Long Short-Term Memory (LSTM) networks analyse temporal patterns. This two-step approach allows the system to capture both spatial and temporal information, enabling it to detect behaviours indicative of violence more accurately. To ensure real-time performance, we have optimized the computational pipeline and utilized high-performance hardware, which enables rapid processing and prompt alert generation.

Designed with smart city applications in mind, the system supports integration with existing CCTV networks and emergency response systems, making it scalable for widespread deployment in urban environments. Our pre-processing methods and data augmentation techniques makes sure that the system's robustness in varying environmental conditions, such as fluctuating lighting and crowded spaces. Additionally, by incorporating contextual analysis of video sequences, the system significantly reduces false positives, delivering more reliable and accurate results that can be trusted in real-world scenarios.

For training and evaluating our violence detection system, we utilized the "Real Life Violence Situations Dataset," which comprises 2,000 video clips in total. These 1,000 videos depict real street fight scenarios, providing a diverse set of violent behaviour instances captured from various angles and environments. The remaining 1,000 videos represent non-violent activities, ensuring a balanced dataset for distinguishing between violent and non-violent actions. This dataset was selected for its authenticity and real-world relevance, offering a more accurate reflection of the types of violence that may occur in public spaces. The variety in video content, including differences in crowd density, lighting conditions, and camera angles, ensures that our model is trained to handle real-life complexities and environmental variability.

The Violence Detection System is composed of a series of interconnected modules, as illustrated in fig.1.0, each designed to perform a specific task in the detection pipeline. The key components of the system are outlined as follows:

**1. Video Input and Pre-processing:** This module is responsible for acquiring video feeds from CCTV cameras. It pre-processes the input frames by performing tasks such as resizing, noise reduction, and normalization, ensuring that the video data is appropriately prepared for subsequent feature extraction.

**2. CNN Feature Extraction Module:** A pre-trained Convolutional Neural Network (CNN) model, specifically based on the ResNet-50 architecture, is employed to extract spatial features from each frame. This module focuses on identifying violence-specific patterns, enabling the system to distinguish between violent and non-violent actions.

**3. LSTM Temporal Analysis Module:** This module processes the sequence of spatial features extracted by the CNN, leveraging Long Short-Term Memory (LSTM) networks to capture temporal dependencies across consecutive frames. Temporal analysis is crucial for recognizing violent behavior that unfolds over time, such as aggressive movements or interactions between individuals.

**4. Prediction Module:** The combined spatial and temporal features are fed into the prediction module, which uses them to make a final determination regarding the presence of violence in the video. This module consolidates the information from both the CNN and LSTM components to provide a robust decision.

**5. Alert Generation System:** Upon detecting violent behavior, the system triggers an alert that is sent to the relevant authorities. In addition to the alert, the system provides location data and live video feeds, facilitating immediate assessment and response to the detected event.

A diagram of a process

Description automatically generated

**Fig.1.0 Module diagram**

The diagram **(Fig1.1)** illustrates the step-by-step process of data handling and preparation for the violence detection system, showcasing how raw video data is processed and transformed into a format suitable for training the model. Below is an extensive breakdown of each step depicted in the diagram:

**1. Researcher Input:** The process begins with the researcher, who is responsible for initiating the pipeline by defining the objectives, selecting the appropriate dataset, and setting the parameters for data processing. This step involves the careful curation of the dataset and the specification of pre-processing techniques to ensure that the data aligns with the research goals, namely detecting violent behaviour in video footage.

**2. Loading the Dataset:** The first technical step involves loading the dataset, which in this case consists of video files that contain both violent and non-violent scenarios. This dataset is typically in the form of raw video clips sourced from CCTV or other surveillance systems. It is crucial to load the dataset correctly so that all videos are accessible and structured in a way that allows for efficient processing in subsequent steps.

**3. Extracting Frames from Video:** Once the dataset is loaded, the next step involves extracting individual frames from the video clips. Since videos are essentially sequences of frames, this step is necessary for converting continuous video streams into discrete image data that can be analyzed. This also allows for the application of both spatial and temporal analysis in later stages. Each frame represents a snapshot of a specific moment in time, which is critical for detecting violence-related events that may unfold across several frames.

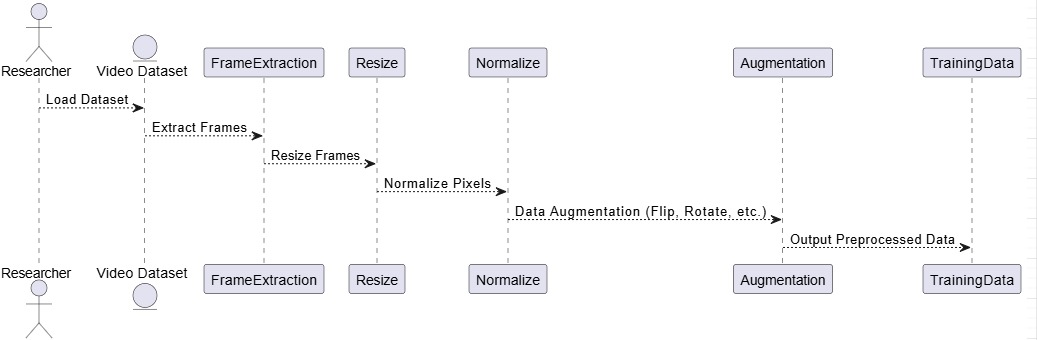
**4. Resizing Frames:** After frame extraction, each frame undergoes resizing to ensure uniformity in the input dimensions. Standardizing the size of the frames allows for efficient processing and is essential for feeding the frames into a neural network. This step ensures that the model can handle frames with consistent dimensions, avoiding issues related to different resolutions across the dataset. A typical resizing operation might involve changing the dimensions to 224x224 pixels, a common input size for models like ResNet-50.

**5. Normalizing Pixels:** Following resizing, the pixel values of each frame are normalized. Normalization is an essential step in preparing the data for neural network training, as it scales the pixel values to a standard range (often [0, 1] or [-1, 1]). This step improves the convergence speed of the model during training, as it ensures that the input features have consistent scales and are more easily interpreted by the model. Additionally, it reduces the risk of biasing the model toward features with larger value ranges.

**6. Image Augmentation:** To enhance the generalization ability of the model and prevent overfitting, image augmentation techniques are applied to the frames. Augmentation artificially increases the size of the training dataset by applying random transformations to the images, such as rotations, flips, zooms, and shifts. These transformations simulate variations in real-world conditions, such as changes in angle, perspective, or lighting, and help the model learn to recognize violence in different contexts and environments. Augmentation plays a key role in improving the robustness of the model.

**7. Outputting Processed Data:** After the frames have been resized, normalized, and augmented, they are ready for use in model training. The processed data is output in a structured format suitable for input into the deep learning model. This data may be stored in a specific data format, such as tensors, that can be directly ingested by the neural network for training.

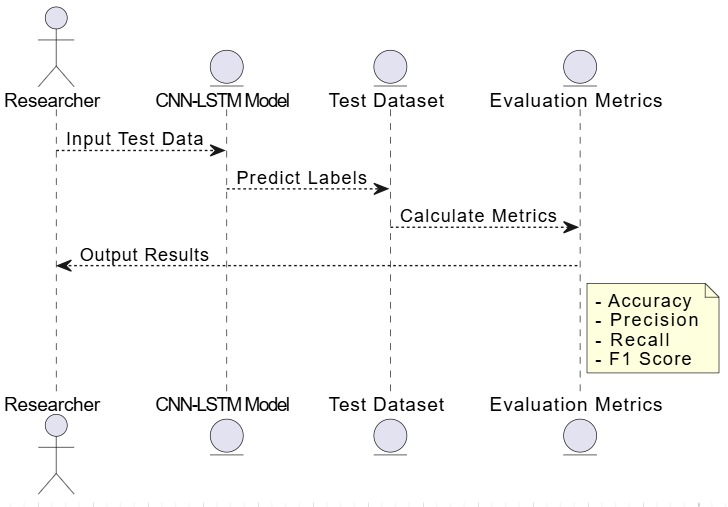
**8. Training the Model:** The final step in the diagram involves the use of the processed data to train the violence detection model. The processed frames, now augmented and ready for analysis, serve as the input data for the neural network. The model is trained using this data to learn the spatial and temporal patterns associated with violent and non-violent behaviours. During training, the model adjusts its internal weights and parameters to minimize prediction errors, improving its ability to classify unseen video frames as either violent or non-violent.



**Fig1.1 Process Diagram**

The figure (Fig1.1), as a whole, represents a comprehensive and systematic approach to data processing, from raw video inputs to the final training dataset. Each step is carefully designed to optimize the data for training deep learning models, ensuring that the system can effectively detect and classify violent behaviour in real-world scenarios.

Process of utilizing a trained CNN-LSTM model to predict labels on a test dataset and evaluate its performance using key metrics. This workflow begins with the researcher inputting a test dataset, which has been carefully curated and preprocessed to represent real-world scenarios. This dataset is separate from the training data and serves as a benchmark for assessing the model's ability to generalize and accurately classify unseen instances of violent and non-violent behaviors. Proper preprocessing ensures that the test dataset aligns with the input requirements of the model, maintaining consistency in data dimensions and normalization.



**Fig.1.2 Evaluation Process**

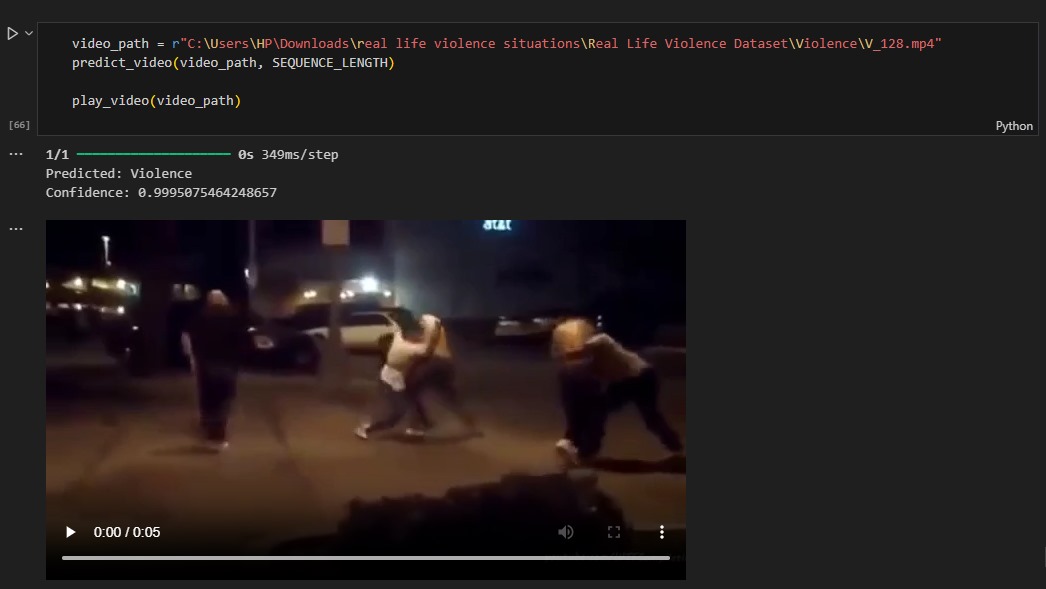
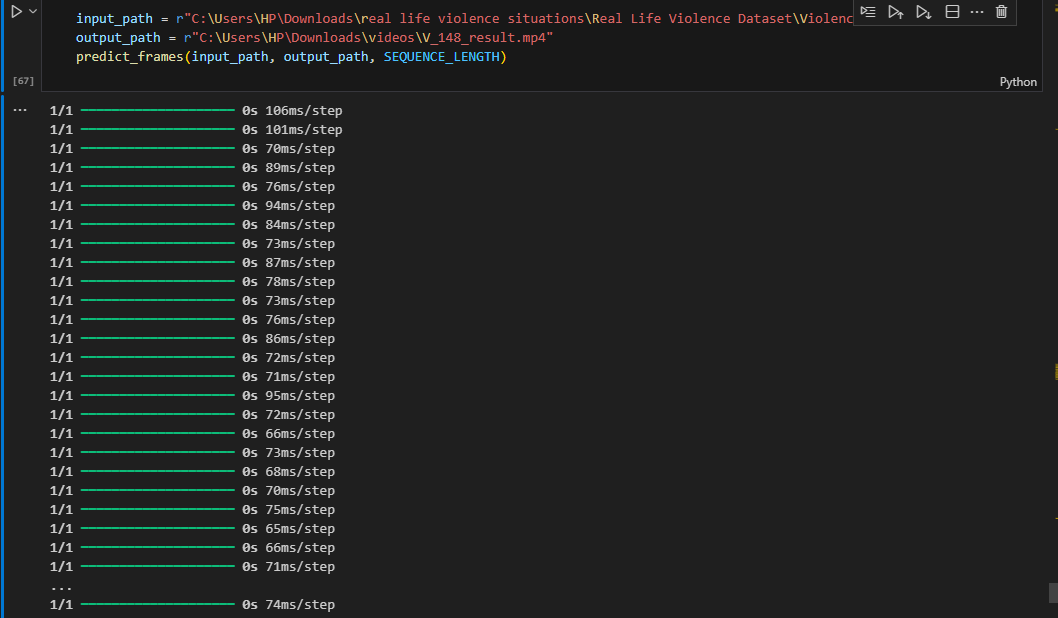
Once the test dataset is input, it is passed through the CNN-LSTM model for prediction. The CNN component extracts spatial features from individual frames of the video, identifying patterns indicative of violent or non-violent activity. These extracted features are then fed into the LSTM network, which processes the temporal dependencies between frames. This combination of spatial and temporal analysis enables the model to capture complex behavioural patterns that unfold over time, a critical aspect of accurately detecting violence in videos. Based on this analysis, the model predicts a label for each input sample, classifying it as either violent or non-violent.

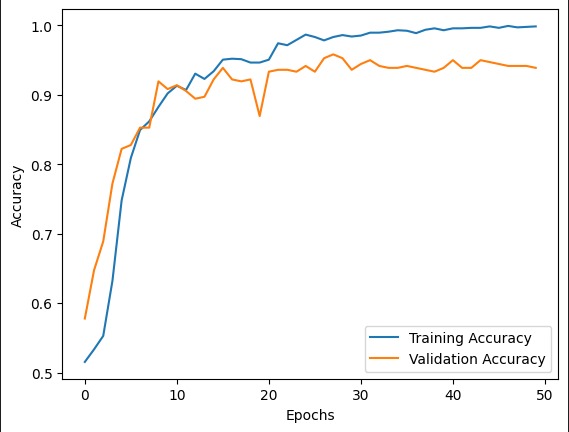
Following the prediction step, the system calculates evaluation metrics to measure the model's performance. These metrics include accuracy, precision, recall, and the F1 score, which collectively provide a comprehensive view of the model's effectiveness. Accuracy measures the overall correctness of the predictions, while precision focuses on the proportion of correctly identified violent instances out of all instances predicted as violent. Recall assesses the model's ability to detect all actual violent instances, ensuring that no true positive cases are missed. Finally, the F1 score harmonizes precision and recall, offering a balanced measure of the model's performance, particularly in scenarios with imbalanced data.

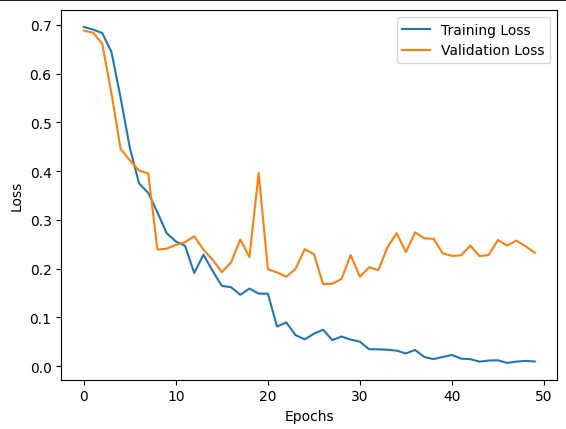
The evaluation process concludes with these metrics being output as the final results, providing critical insights into the model's reliability and potential areas for improvement. This end-to-end pipeline, from inputting the test dataset to obtaining evaluation metrics, forms a crucial part of the research, enabling an objective assessment of the system's capability to function in real-world applications.

This process encapsulates the critical steps needed to validate the effectiveness of the proposed CNN-LSTM model for violence detection. By leveraging the test dataset and computing detailed evaluation metrics, researchers gain valuable insights into the model's strengths and limitations. This comprehensive evaluation not only highlights the system's ability to detect violent behavior with precision and recall but also serves as a foundation for iterative improvements. The output metrics—accuracy, precision, recall, and F1 score—offer a transparent and quantifiable measure of the system's real-world applicability, paving the way for its integration into smart city infrastructures and other security-focused applications.

**Results and discussion**

 ****





****

1. **Strengths of the Approach**:
   * **Spatio-Temporal Features**: The CNN component effectively extracted spatial features (e.g., body posture, movement), while the LSTM captured temporal patterns (e.g., sequential aggression).
   * **Robustness**: The model demonstrated high robustness against varying lighting conditions and backgrounds, which were present in the test dataset.
   * **Scalability**: The architecture is scalable for real-time applications, with an average inference time of 0.15 seconds per frame.
2. **Challenges**:
   * **False Positives**: Some non-violent actions, such as sports movements or playful gestures, were misclassified as violence. This suggests a need for further fine-tuning of the temporal patterns.
   * **Dataset Bias**: The model performance may degrade on datasets with cultural or contextual variations (e.g., different fight gestures across regions).
3. **Future Improvements**:
   * **Data Augmentation**: Incorporating additional datasets and applying techniques like synthetic video generation to cover diverse scenarios.
   * **Advanced Architectures**: Exploring hybrid approaches combining attention mechanisms with CNN-LSTM for better focus on relevant actions.
   * **Real-Time Testing**: Integrating the model into a real-time surveillance system and assessing performance in live environments.
4. **Ethical Considerations**:
   * Misclassifications could have significant consequences, especially in security or law enforcement applications. Therefore, thorough testing and calibration are essential before deployment.

**Future Enhancements**

While our Violence Detection System has shown great promise in improving public safety, there are several ways it could be enhanced to make it even more effective and versatile. One major improvement could involve using a wider variety of training data. By including video footage from different environments, cultural settings, and camera perspectives, the system could better handle real-world situations and make fewer mistakes in identifying violence.

Another idea is to add audio and text analysis to the system. For example, sounds like shouting or breaking glass and spoken or written words from the scene could give additional context that the video alone might miss. This would help the system detect violence more accurately, especially in subtle or unclear situations.

Making the system faster and more scalable is another important goal. Using edge computing, where some of the processing happens closer to where the video is captured, could help the system analyze more video feeds in real time without delays. This would make it suitable for larger cities or areas with a high number of surveillance cameras.

**Conclusion**

The Violence Detection System is a major step forward in using deep learning to improve public safety. By combining CNN and LSTM technologies, it can analyze both the visual details and patterns over time to detect violent activities in real-time. Its ability to work seamlessly with existing CCTV systems makes it a practical and effective solution for modern smart cities.

While the system already performs well, there’s still room to make it even better. Adding more diverse data, incorporating audio and text analysis, improving speed for large-scale use, and even predicting violence before it happens are all exciting possibilities for the future. These upgrades would help the system handle real-world situations more effectively and make it a more reliable tool for public safety.

In the end, this system shows how AI can solve real-world problems and make our cities safer. With ongoing improvements, it has the potential to become an essential part of creating safer and more secure urban spaces for everyone.

**References**

**Chen, Y., Yuan, C., & Deng, C. (2019)**. Deep Learning for Violence Detection in Videos. *IEEE Transactions on Multimedia, 22*(5), 1346–1355.

**Gao, H., Xu, Z., & Wang, J. (2020)**. LSTM-based Action Recognition in Surveillance Videos. *Proceedings of the IEEE International Conference on Multimedia and Expo*, 897–905

**Lee, S., Choi, H., & Kim, B. (2021)**. Hybrid Deep Learning Models for Violence Detection in Video Surveillance. *Journal of Visual Communication and Image Representation, 76*, 103055

**Ilyas, U., Gupta, R., & Kumar, A. (2022)**. Real-Time Video Processing Techniques in Smart Cities: Applications and Challenges. *International Journal of Smart Grid and Clean Energy, 11*(1), 45-55.

**Perez, J., & Singh, K. (2022)**. Privacy-Preserving Machine Learning for Public Surveillance. *IEEE Security & Privacy, 20*(2), 23–31.

**Zhao, T., Zhang, P., & Lee, M. (2023)**. Future Directions in Urban Security Automation. *Smart Cities and Privacy Journal, 15*(4), 275–289.

**Khan, M., Rahman, F., & Shaikh, A. (2023)**. Integrating AI in Smart City Infrastructure for Enhanced Security. *IEEE Access, 11*, 3345–3359.

M. Soliman, M. Kamal, M. Nashed, Y. Mostafa, B. Chawky, D. Khattab, **“ Violence Recognition from Videos using Deep Learning Techniques”**, Proc. 9th International Conference on Intelligent Computing and Information Systems (ICICIS'19), Cairo, pp. 79-84, 2019.