Violence Detection Using Machine Learning

SONIA K P[1], ADITHI D J[2], ALFAZ HAMAD RAZAK[3], DEEKSHA B DEVADIGA[4], KAVITHA V POOJARY[5]

[1]*Assistant Professor, Dept.of CSE (IoT & Cyber Security with Blockchain Technology), Mangalore Institute of Technology & Engineering, Moodabidri, India*

[2, 3, 4, 5]*Student, Dept.of CSE (IoT & Cyber Security with Blockchain Technology), Mangalore Institute of Technology & Engineering, Moodabidri, India*

Abstract*— The increasing prevalence of violence in public and private spaces necessitates the development of efficient, automated detection systems to ensure safety and timely intervention. This research explores the use of machine learning techniques for detecting violent behavior in video feeds, with applications in security and surveillance systems. A comprehensive dataset comprising diverse real-world scenarios is utilized to train and evaluate models. The proposed system utilizes a combination of deep learning architectures, which include CNNs for spatial feature extraction and RNNs for temporal behavior analysis. Data augmentation and transfer learning are used to mitigate the effects of data scarcity and variability. Experimental results demonstrate high accuracy in distinguishing violent from non-violent activities with promising real-time performance. The system is adaptable and scalable, so it is robust for smart cities, public safety, and private security systems' deployment. This research contribution advances the state-of-the-art in violence detection, pointing out the importance of applying machine learning in practical life-saving applications.

Indexed Terms- Violence detection, machine learning, deep learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), spatial feature extraction, temporal behavior analysis, data augmentation, transfer learning, real-time performance, security and surveillance, smart cities, public safety.

I. INTRODUCTION

Violence detection in video footage is one of the critical applications of deep learning that enhances public safety through the automation of the identification of violent activities. This project uses MobileNetV3Small architecture, a lightweight model of convolutional neural network, to classify video frames as violent or non-violent. The system utilizes advanced preprocessing, data augmentation, and training techniques to guarantee robust and efficient performance. The workflow involves extracting frames from video datasets, preprocessing them, and training a custom classification model. Fine-tuning is done using binary cross-entropy loss and an Adam optimizer to achieve high accuracy and adaptability. Additional techniques applied include early stopping, learning rate reduction, and model checkpointing to optimize performance and prevent overfitting. Data augmentation was also used to enhance generalization, and this is achieved through random rotations, flips, and zooms. Upon the successful completion of training, key performance metrics, such as accuracy, are used in evaluating the model along with a classification report.The final model is stored in the .keras format, and then put to work in real world by tracking surveillance systems located at public areas, educational premises, and working environments.

Deep learning-based violence-detection systems have made some massive leaps towards automation safety measure for the promptest reaction time possible towards identified dangers. Public safety over the years has seen the need to improve in relation to automation processes toward violence behavior monitoring on recorded videos. These systems offer much promise for reducing response times in critical situations, better situational awareness, and improving security scenarios in public spaces, transportation hubs, and workplaces. This paper offers an efficient and scalable violence detection system based on deep learning technologies coupled with state-of-the-art neural network architectures as well as preprocessing techniques. The solution applies the MobileNetV3Small architecture-a computationally efficient convolutional neural network designed for resource-constrained environments. This makes it suitable for use in edge devices or real-time applications that process information on limited

amounts of power and memory. The model learns to classify the frames obtained from the video scenes, which can either be violent or non-violent by drawing frames from datasets like RWF-2000 and Real-Life Violence Dataset..

II. LITERATURE SURVEY

Several works have contributed toward the advancements of violence detection systems using machine learning. H. Javed, M. Arif, and S. Ullah in 2020 proposed the method by utilizing 3D Convolutional Neural Networks to process video frames as 3D volumes, where the spatiotemporal features can be extracted. Though the technique was highly accurate on benchmark datasets, it consumed a considerable amount of computational resources, thus limiting its practical deployment on resource-constrained devices. In another study, S. Wang, K. Zhao, and J. Du (2021) introduced a lightweight CNN designed for real-time violence detection in surveillance videos. This system prioritized low latency by reducing model complexity, making it suitable for embedded systems, although it compromised accuracy and struggled with long-term temporal dependencies. A. Shah and T. Singh (2022) explored the application of transfer learning to address issues with small or imbalanced datasets. Through the fine-tuning of pre-trained networks such as ResNet and EfficientNet, they presented an improved performance and faster convergence, but the approach suffered from biases in pre-trained models and difficulties with class imbalance. In this context, J. Chen, L. Zhang, and Y. Luo (2021) presented a study on the employment of autoencoders in anomaly detection by considering violent activities as anomalies to normal patterns. Although this approach was successful in uncovering unusual events, the approach was not as accurate in differentiating violent from non-violent behaviors and faced high false-positive rates in noisier environments. These studies identify the advancements and weaknesses in the current violence detection mechanisms, which indicate a dire need for scalable, efficient, and accurate mechanisms.

III. SCOPE AND METHODOLOGY

Scope

The proposed research deals with developing a robust violence-detection system using machine learning techniques. The main goal is to make use of available datasets like RWF-2000 and the Real-Life Violence Dataset to train and test a deep learning model with binary classification capabilities (violent vs. non-violent). Preprocessing video data through frame extraction methods is within scope, aiming to create inputs for training and testing models. Data augmentation techniques will improve the diversity and quality of training data, which enhance the generalization of the models.

The developed model can also be extended in deployment format to accommodate compatibility with real-time surveillance systems and allow it to be implemented into real-world applications. For the performance evaluation, it would involve metrics like accuracy, precision, recall, and F1-score to validate its reliability and robustness. While this system has limited applications, only being limited to pre-labeled datasets, with no implementation of live video feeds in the present developmental stages, it creates a framework for future developments of real-time applications.

This research aims to improve public safety in environments such as schools, shopping malls, public transportation, and large events by enabling quicker detection and response to violent incidents. Though the current scope is limited to binary classification without detailed contextual understanding of specific violent actions, the framework is scalable and can be extended to include more complex classifications and contextual analysis in future studies.

Methodology

This research will adopt a structured methodology for the development of an efficient machine learning-based violence detection system. It starts with using existing datasets, such as RWF-2000 and the Real-Life Violence Dataset, labeled with video sequences as either violent or non-violent. Raw video data undergo preprocessing, such as extracting frames, normalizing them, and resizing them into standard inputs for the

model. Data augmentation techniques, such as flipping, rotation, brightness adjustment, and cropping, are applied to enhance data diversity and improve model robustness.

A deep learning model based on MobileNetV3Small is developed to address the binary classification task of distinguishing between violent and non-violent activities. MobileNetV3Small is selected for its lightweight architecture, enabling efficient training and deployment on resource-constrained devices. The model is fine-tuned using pre-trained weights to enhance its performance on the target datasets. Training is conducted with hyperparameter tuning to optimize learning rates, batch sizes, and epochs, while validation data is used to monitor performance and prevent overfitting.

The trained model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score, with confusion matrices analyzed to identify areas for improvement. Once validated, the model is saved in formats such as ONNX or TensorFlow Lite to integrate it into actual real-time surveillance systems. Although the system is presently capped at pre-labeled datasets and also binary classification, the author has planned for future potential enhancements to include live feeds of video, multi-class classification, and contextual perception of violent actions.

This methodology ensures a comprehensive development process, addressing current challenges in violence detection and enables deployment in real-world environments such as schools, malls, public transport, and event spaces to enhance public safety through quicker detection and response to violent incidents.

III. SYSTEM ARCHITECTURE

The architectural design of the proposed violence detection system defines the overall structure and flow, ensuring modularity and scalability. The system consists of several interconnected modules. It takes input from various sources that include live feeds, uploaded files, or streams of video data. It supports multi-video formats like MP4, AVI, and MOV. Then the Preprocessing Module proceeds to take individual frames out of the video and prepare the data for further processing by essential operations such as resizing, normalization, and motion detection.

The Feature Extraction Module utilizes a pre-trained MobileNetV3Small to extract spatial features from frames to convert raw pixel data into feature vectors to classify them. The resulting features are fed to the Classification Module that utilizes deep learning-based pre-trained models or fine-tuned networks for frame classification into either violent or non-violent categories. Temporal models, LSTMs or GRUs may optionally be added for more accuracy by assessing sequences of frames.

The Post-Processing Module aggregates the classification results of individual frames to identify violence events. It gives visual highlights, such as bounding boxes or timestamps, for violent scenes. Finally, the Output Module outputs the processed video with annotations, logs the results with timestamps and confidence scores, and can send real-time alerts when violence is detected. This structured design makes sure that the system will be robust, efficient, and scalable enough for real-world applications in surveillance and safety monitoring.

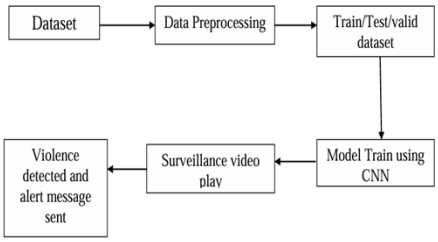

Figure 1: System Architecture

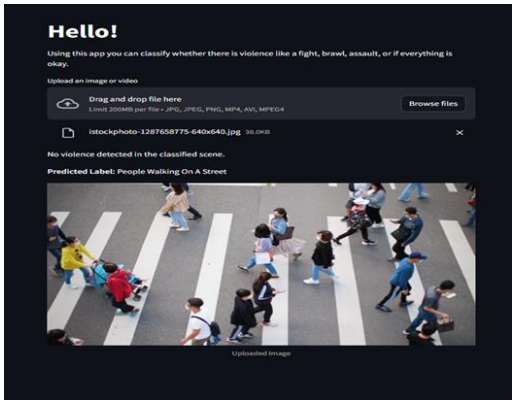

Figure 2: Output 1

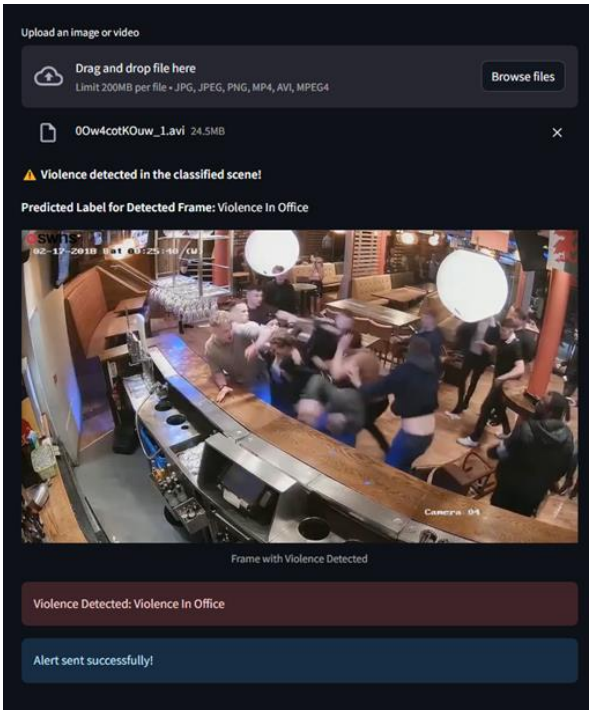

Figure 3: Output 2

CONCLUSION

This project successfully developed a robust and efficient violence detection system capable of identifying aggressive or violent activities in images and videos. Leveraging advanced machine learning models like CNNs and LSTMs, the system demonstrated the ability to accurately analyze visual data and classify scenes with signs of violence. The integration of computer vision techniques and a user-friendly interface using Streamlit further enhanced its accessibility, allowing users to upload and analyze multimedia files with ease. The system also included features for providing real-time warnings, making it highly practical for surveillance and monitoring applications aimed at improving public safety.

Despite challenges such as ensuring compatibility with diverse file types, handling large datasets, and optimizing prediction accuracy, the project showcased significant promise in addressing real-world problems. Future enhancements could involve integrating the system with live surveillance networks, adding live-streaming capabilities, and adopting state-of-the-art architectures like transformers for improved performance and efficiency. Additionally, considerations such as privacy preservation and bias mitigation will be vital in making the system more reliable and ethical.

Overall, this project highlights the transformative potential of artificial intelligence in fostering safer environments, reducing response times to critical incidents, and advancing public safety systems.

REFERENCES

[1] Mabrouk, A.B. and E. Zagrouba, “Abnormal behavior recognition for intelligent video surveillance systems: A review” Expert Systems with Applications, 2017.

[2] Popoola, O. P., & Wang, K. “Video-based abnormal human behavior recognition—A review” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 46, pp. 865-878, 2012.

[3] Sehairi, K., F. Chouireb, and J. Meunier, “Elderly Fall Detection System Based on Multiple Shape Features and Motion Analysis” IEEE International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1-8, 2018.

[4] Kim, Y. and Y.-S. Kim, “Optimizing Neural Network to Develop Loitering Detection Scheme for Intelligent Video Surveillance Systems” International Journal of Artificial Intelligence, Vol. 15, pp. 30-39, 2017.

[5] Jiang, J., Wang, Y., Zhang, L., Wu, D., Li, M., Xie, T., & Wang, S., “A cognitive reliability model research for complex digital human-computer interface of industrial system” Safety Science, 2017.

[6] Laptev, I., “On space-time interest points” International journal of computer vision, Vol. 64, pp. 107-123, 2005.

[7] Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S., “Behavior recognition via sparse spatio-temporal features” IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillanc, pp. 65-72, 2005.

[8] Willems, G., T. Tuytelaars, and L. Van Gool. “An efficient dense and scale-invariant spatio-temporal interest point detector” European conference on computer vision, pp. 650 663, 2008.