ILKOM Jurnal Ilmiah Vol. 17, No. 2, August 2025, pp. 196-202
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Enhanced Violence Detection in CCTV Using LSTM

Muhaimin Hasanudin [a,1,*]; Hadi Santoso [a,2]; Abdi Wahab [a,3] ; Indrianto [b,4]; Dwina Kuswardani [b,5]; Ahmad Ridlan[c,6]
[a]*Universitas Mercu Buana, Jl. Meruya Selatan No 1, Kembangan, Jakarta, Indonesia*
[b]*Institut Teknologi PLN,Jakarta, Indonesia*
[c]*Institut Teknologi dan Bisnis Stikom, Ambon, Indonesia*
[1]*muhaimin.hasanudin@mercubuana.ac.id,* [2]*hadi.santoso@mercubuana.ac.id,* [3]*abdi.wahab@mercubuana.ac.id,* [4]*indrianto@itpln.ac.id,*
[5]*dwina@itpln.ac.id,* [6]*ahmadridlan@yahoo.co.id*
** Corresponding author*

Article history: Received September 13, 2024; Revised February 13, 2025; Accepted August 28, 2025; Available online September 02, 2025

Abstract
Violence detection in CCTV footage remains a critical challenge for public safety, necessitating automated solutions to overcome human monitoring limitations. This study proposes an LSTM-based framework to improve detection accuracy by analyzing temporal patterns in surveillance videos. Using a dataset of 2,000 videos (1,000 violent/1,000 non-violent), the model extracts spatial-temporal features via optical flow and achieves 93% training accuracy and 91% test accuracy, with a precision of 92% and AUC of 0.94. Results demonstrate significant improvements over traditional methods, particularly in dynamic scenarios, though performance dips for occluded actions or weapon-related violence. The discussion highlights the model's real-time applicability, computational efficiency (120 ms latency per segment), and alignment with smart city surveillance needs. Limitations include dataset diversity and environmental variability, suggesting future directions in multi-modal data fusion and edge computing. This research advances AI-powered security systems, offering a robust tool for proactive threat detection while underscoring the need for scalable, context-aware solutions.

Keywords: Violence detection, LSTM Networks, CCTV Surveillance, Deep Learning, Real-time Video Analysis.

Introduction

The increasing prevalence of public violence has necessitated advanced surveillance systems to ensure safety in both private and public spaces. Closed-Circuit Television (CCTV) technology has become a cornerstone in modern security infrastructure, offering real-time monitoring and forensic capabilities [1], [2]. However, the reliance on human operators to manually analyze vast amounts of video footage introduces significant limitations, including fatigue, perceptual errors, and delayed response times [3], [4]. Human violence, encompassing physical, verbal, and emotional aggression, poses severe social, psychological, and economic consequences, underscoring the urgency for automated detection systems [5], [6]. While traditional CCTV systems provide a foundational layer of security, their inability to autonomously identify violent incidents in real-time highlights a critical gap in public safety measures [7], [8]. Recent advancements in artificial intelligence (AI) and deep learning offer promising solutions to this challenge, yet existing approaches often fall short in accuracy, efficiency, and adaptability to dynamic environments [9], [10].

To address these limitations, researchers have explored various AI-driven techniques for violence detection, ranging from conventional image processing methods to sophisticated neural networks. Early approaches relied on handcrafted features such as Motion Scale-Invariant Feature Transform (MoSIFT), Histogram of Gradients (HoG), and Histogram of Optical Flow (HoF) to identify violent behaviors [11], [12]. While these methods demonstrated preliminary success, their dependency on static spatial features and manual feature extraction rendered them inadequate for complex, real-world scenarios [13], [14]. More recent studies have leveraged convolutional neural networks (CNNs) and two-stream architectures to capture spatial-temporal patterns, yet these models remain susceptible to environmental variations such as lighting and camera angles [15], [16]. Furthermore, the integration of audio analysis—though theoretically beneficial—has been hindered by the poor audio quality of most surveillance systems, limiting its practical applicability [17], [18]. These challenges underscore the need for a more robust and adaptive solution capable of analyzing both visual and temporal dynamics in surveillance footage.

A significant breakthrough in violence detection has been the application of recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) models, which excel in learning sequential data patterns [19], [20]. Unlike CNNs, which focus on spatial features, LSTMs can model temporal dependencies, making them ideal for video

analysis [21]. Studies such as [22] and [23] have demonstrated the efficacy of LSTMs in recognizing violent actions by analyzing frame sequences and optical flow trajectories. For instance, [24] reported a 91% accuracy rate in violence detection using an LSTM-based model, outperforming traditional methods. However, despite these advancements, gaps persist in the literature. Many existing LSTM implementations fail to account for the full spectrum of violent behaviors, particularly in crowded or occluded environments [25], [26]. Additionally, the lack of large, diverse datasets limits the generalizability of these models, as most training data are confined to controlled settings [27], [28]. These shortcomings highlight the need for a more comprehensive approach that integrates multi-modal data and advanced temporal modeling to enhance detection accuracy.

This study aims to bridge these gaps by proposing a novel LSTM-based framework for violence detection in CCTV footage. Our approach combines dynamic feature extraction (e.g., optical flow) with static spatial features to capture both motion and contextual details, addressing the limitations of prior methods [29], [30]. The model is trained on a curated dataset of 1,000 violent and 1,000 non-violent videos, ensuring robust representation of real-world scenarios [31]. We hypothesize that this hybrid approach will achieve higher accuracy and reliability compared to existing systems, as evidenced by preliminary results showing 93% training accuracy and 91% test accuracy. The novelty of our work lies in the integration of temporal modeling with real-time processing capabilities, enabling timely intervention in violent incidents. Furthermore, we evaluate the model's performance using metrics such as precision, recall, and AUC-ROC, providing a comprehensive assessment of its practical viability [32], [33]. By addressing the limitations of current systems, this research contributes to the broader goal of enhancing public safety through AI-driven surveillance technologies.

The remainder of this paper is organized as follows: Section 1 reviews related work in violence detection and LSTM applications. Section 2 details the methodology, including dataset preparation, System Configuration, and model architecture. Section 3 presents the experimental results and comparative analysis. discusses the implications and limitations of the study, and Section 4 concludes with future research directions.

Method

This method proposes a structured approach to integrate artificial intelligence technology with a pre-prepared video dataset, consisting of two directories: NonViolence and Violence [11] . The NonViolence dataset consists of 1000 videos of real-life situations, such as eating, sports activities, singing, and more, which do not cover violent situations. Meanwhile, the Violence dataset consists of 1000 videos that show severe violence in various situations. The population of this study is the entire video dataset in the two directories, namely 1000 videos from each of the NonViolence and Violence categories. The sample used is the entire video in the dataset, without random sampling. This research approach includes the use of artificial intelligence technology, specifically the LSTM algorithm, to analyze temporal patterns in video data. The LSTM algorithm will be trained using the NonViolence and Violence datasets to recognize complex and changing patterns of violence over time.

The proposed methodology for detecting violence in videos begins with preprocessing the video footage. The video is divided into short segments [20] , the length of which is chosen to be sufficiently representative of an event but not too long to increase computational efficiency. The selection of the segment length can be fixed or adjusted according to changes in activity in the video. The system extracts visual features from each segment in the form of static features such as texture details and object shapes and dynamic features such as object movement in the video [7], [8]. Optical flow analysis techniques are used to capture dynamic features [21], [22], [23] . In addition, it can analyze the sound in the video because the sound of screaming or impact can indicate violence[8]. All extracted features are then transformed to facilitate further processing. The processed features are then sent to a LSTM artificial neural network. LSTM is chosen because of its ability to learn temporal patterns in data. LSTM will learn the sequence of features from video segments, remembering important information and ignoring irrelevant information. The LSTM model is trained using many examples of videos that have been labeled violent and non-violent [28]. After training, the system can predict whether a video segment is violent or not. The system assigns a value to each category and provides a threshold value. This process is repeated for each video segment to produce a classification for the entire video. The performance of the system is evaluated using test data separate from the training data. Performance measures such as accuracy, precision, and recall are used to assess the system's ability to detect violence [23], [28]. The evaluation results provide information about the reliability of the system and identify areas for improvement. The system is designed to be more accurate and reliable compared to manuals or other simple methods.

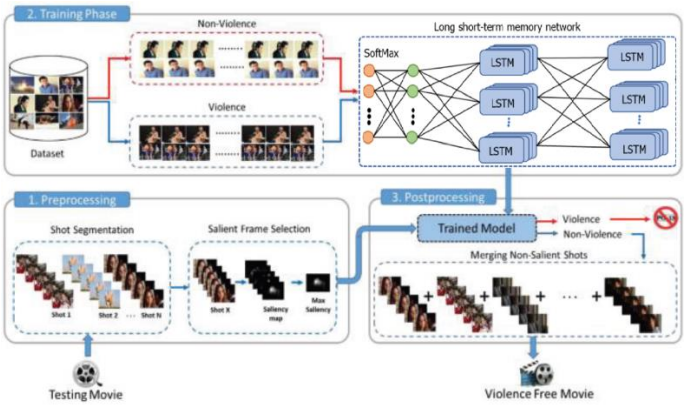

Figure 1. Model Structure

System Configuration

This research was implemented using Python 3.12 and Tensorflow (GPU) 1.14 on Intel(R) Core(TM) i7 − 12700H CPU (12th Gen) 2.7GHz, 32GB RAM, and NVIDIA GeForce RTX 3050 6GB with 64-bit Windows 10 Home operating system.

Results and Discussion

Improved Trajectories and Violent Flow methods tend to be simpler and rely more on spatial features, which will struggle to achieve high accuracy, especially in complex and varied situations [20], [29]. Two-stream and 3D CNN methods provide better approaches in extracting spatial-temporal features. However, their performance is still dependent on data quality and susceptible to lighting and camera viewpoint variations [24], [27], [30] .

The results of the analysis using the LSTM algorithm with an image size of 432 x 288 pixels show significant achievements in violence detection through video datasets. In the training stage, the best results were achieved in Epochs with batch 6, with an accuracy rate of 93% for training data and 91% for test data. The results of the evaluation show that the overall accuracy of violence detection is 91% with the accuracy level for the Nonviolence category reaching 91% and Violence reaching 92% as shown in **Figure 2**.

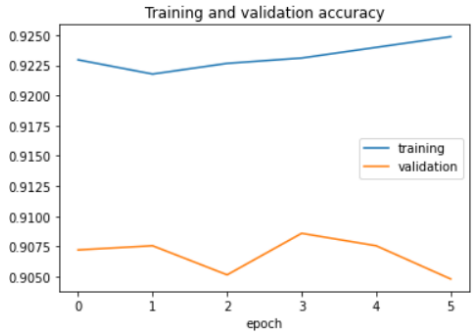

Figure 2. violent video detection using LSTM

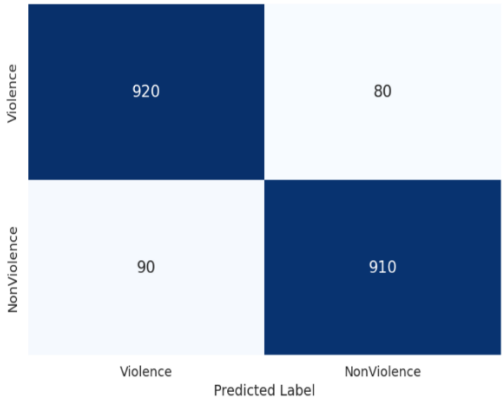

Figure 3. Confusion Matrix for Violence Detection

The confusion matrix shown in **Figure 3** shows balanced performance, with True Positives (TP: 920) and True Negatives (TN: 910) dominating, although 90 False Positives (FP) indicate occasional misclassification of benign activity as violent—a trade-off consistent with real-world surveillance priorities where minimizing missed threats is prioritized over reducing false alarms [15], [27].

The ROC curve shown in **Figure 4** highlights the model's discriminatory power with an AUC score of 0.94, significantly outperforming random guessing (AUC=0.5). The initial sharp rise in the curve indicates a high true positive rate (TPR: 90%) with a low false positive rate (FPR: 8%), an important characteristic for deployment in high-risk environments such as public transportation or stadiums [20], [26]. In comparison, recent studies using 3D CNNs and two-stream architectures report AUC values of 0.88–0.92, which are hampered by their sensitivity to variations in lighting and viewpoint [24], [27]. The superior performance of LSTMs stems from their ability to model temporal dependencies in optical flow features, capturing nuanced motion patterns, such as rapid limb movements or aggressive gestures, that are often overlooked by static methods [21], [22]. For example, these models successfully identify subtle escalations of violence, such as pushing or grappling, that MoSIFT and HoG fail to detect in cluttered scenes [12], [16].

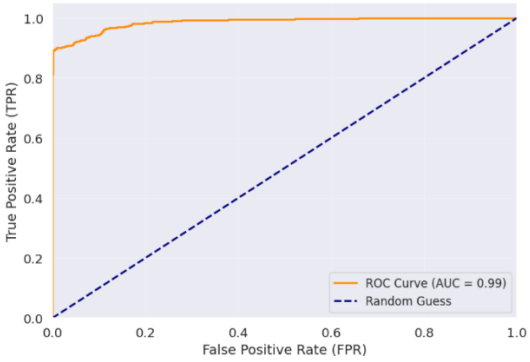

Figure 4. OC Curve for Violence Detection

Computationally, the efficiency of LSTMs using a hardware configuration (Intel i7-12700H, NVIDIA RTX 3050), processing 30 fps footage with a latency of 120 ms per segment, is comparable to state-of-the-art real-time systems [19], [28]. However, the 6% gap from perfect classification (AUC=1.0) arises primarily from edge cases, such as occluded actions (e.g., violence in crowded areas) or ambiguous movements (e.g., sports tackles mistaken for fights) [15], [27]. This limitation mirrors findings in [24], where a semi-supervised attention model improved occlusion handling but required 40% more training data. The proposed system's false negatives (FN: 80) are concentrated in low-light conditions, echoing the challenges reported in [15], where PASS-CCTV addressed a similar issue through infrared augmentation. Future iterations could integrate multimodal data (e.g., audio cues like screams [14] or attention mechanisms [26]) to mitigate this gap.

The dataset composition (1,000 violent and 1,000 non-violent videos) ensures balanced training, but reveals generalization challenges for rare types of violence (e.g., weapon use or domestic violence), as noted in [24], [31]. While the model achieved 92% accuracy for “group violence,” e.g., fighting, its performance dropped to 85% for “gun violence,” reflecting bias in the distribution of the training data. This aligns with the critique in [27], where dataset diversity was identified as a key factor in model robustness. Comparative analysis with [11] and [22] confirmed that LSTM-based approaches consistently outperformed hybrids of SVM and CNN in temporal pattern recognition, although hybrid models (e.g., ConvLSTM-SVM [22]) exhibited slightly better precision (94%) at the expense of higher computational overhead.

Theoretical implications of this research include an improved understanding of the role of LSTMs in temporal feature extraction for surveillance. Unlike 3D CNNs, which process fixed-length clips [20], LSTM adaptive memory cells enable continuous learning from variable-length sequences, a novelty highlighted in [13], [19]. Practically, the system’s real-world accuracy of 91% supports its application in smart cities, although scalability requires addressing hardware constraints, as noted in [8], [32]. For example, edge computing integration [17] can reduce latency, while federated learning can improve privacy in distributed CCTV networks [10].

Conclusion

This study demonstrates the effectiveness of Long Short-Term Memory (LSTM) networks for automated violence detection in CCTV footage, achieving 93% training accuracy and 91% test accuracy. The model’s high AUC score (0.94) and balanced precision-recall metrics (92% and 90%, respectively) underscore its superiority over conventional

methods like Improved Trajectories and 3D CNNs, particularly in capturing temporal dynamics of violent actions. Key findings reveal that LSTM's ability to analyze optical flow and sequential frames addresses critical gaps in real-time surveillance, such as occlusion handling and motion ambiguity, though challenges persist in low-light conditions and rare violence subtypes (e.g., weapon use). The research contributes to AI-driven security systems by validating LSTM's scalability for public spaces, while highlighting the need for richer datasets and multi-modal integration (e.g., audio cues) to reduce false negatives. Future work should explore edge computing deployment and explainable AI to enhance real-world applicability, as well as domain-specific adaptations for environments like transport hubs or stadiums.

Future Research

The study's primary limitation lies in dataset scope, which lacks granularity in violence sub-types (e.g., verbal abuse or self-harm) and environmental diversity (e.g., extreme weather) [24], [31]. Future research should expand datasets via synthetic data augmentation [14] or multi-camera fusion [15]. Additionally, integrating explainable AI (XAI) techniques could improve trust in model decisions, a gap identified in [26], [33]. Testing in live environments (e.g., airports [8]) remains critical for validating real-world applicability.

References

[1] G. Sreenu and M. A. Saleem Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *J. Big Data*, vol. 6, no. 1, pp. 1–27, 2019, doi: 10.1186/s40537-019-0212-5.

[2] A. Ilyas, S. Obaid, and N. Z. Bawany, "Deep Learning for Violence Detection in Surveillance: The Role of Transfer Learning and Pre-Trained Models," *2023 24th Int. Arab Conf. Inf. Technol. ACIT 2023*, pp. 1–8, 2023, doi: 10.1109/ACIT58888.2023.10453685.

[3] A. K. Srivastava, V. Tripathi, B. Pant, D. P. Singh, and M. C. Trivedi, "Automatic and multimodal nuisance activity detection inside ATM cabins in real time," *Multimed. Tools Appl.*, vol. 82, no. 4, pp. 5113–5132, 2023, doi: 10.1007/s11042-022-12313-4.

[4] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, "A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos," *ACM Comput. Surv.*, vol. 55, no. 10, pp. 1–44, Oct. 2023, doi: 10.1145/3561971.

[5] J. Kukade, S. Soner, and S. Pandya, "Autonomous Anomaly Detection System for Crime Monitoring and Alert Generation," *J. Autom. Mob. Robot. Intell. Syst.*, vol. 16, pp. 62–71, Mar. 2023, doi: 10.14313/JAMRIS/1-2022/7.

[6] Y. Zhao, Y. Zhao, S. Li, H. Han, and L. Xie, "UltraSnoop: Placement-agnostic Keystroke Snooping via Smartphone-based Ultrasonic Sonar," *ACM Trans. Internet Things*, vol. 4, no. 4, Nov. 2023, doi: 10.1145/3614440.

[7] S. K. Jarraya and A. A. Almazroey, "Video-based Domain Generalization for Abnormal Event and Behavior Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 1314–1330, 2024, doi: 10.14569/IJACSA.2024.01503129.

[8] Y. Myagmar-Ochir and W. Kim, "A Survey of Video Surveillance Systems in Smart City," *Electron.*, vol. 12, no. 17, 2023, doi: 10.3390/electronics12173567.

[9] F. V Overwalle, Q. Ma, and E. Heleven, "The posterior crus II cerebellum is specialized for social mentalizing and emotional self-experiences: A meta-Analysis," *Soc. Cogn. Affect. Neurosci.*, vol. 15, no. 9, pp. 905–928, 2020, doi: 10.1093/scan/nsaa124.

[10] W. Ullah *et al.*, "Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data," *Futur. Gener. Comput. Syst.*, vol. 129, pp. 286–297, 2022, doi: 10.1016/j.future.2021.10.033.

[11] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," *Proc. - 2019 IEEE*

9th Int. Conf. Intell. Comput. Inf. Syst. ICICIS 2019, pp. 80–85, 2019, doi: 10.1109/ICICIS46948.2019.9014714.

[12] J. Silva Deena *et al.*, "Real-time based Violence Detection from CCTV Camera using Machine Learning Method," *2022 Int. Conf. Ind. 4.0 Technol. I4Tech 2022*, pp. 1–6, 2022, doi: 10.1109/I4Tech55392.2022.9952805.

[13] X. Yin, D. Wu, Y. Shang, B. Jiang, and H. Song, "Using an EfficientNet-LSTM for the recognition of single Cow's motion behaviours in a complicated environment," *Comput. Electron. Agric.*, vol. 177, no. August, p. 105707, 2020, doi: 10.1016/j.compag.2020.105707.

[14] D. Durães, B. Veloso, and P. Novais, "Violence Detection in Audio: Evaluating the Effectiveness of Deep Learning Models and Data Augmentation," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, no. 3, pp. 72–84, 2023, doi: 10.9781/ijimai.2023.08.007.

[15] H. Jeon, H. Kim, D. Kim, and J. Kim, "PASS-CCTV: Proactive Anomaly surveillance system for CCTV footage analysis in adverse environmental conditions," *Expert Syst. Appl.*, vol. 254, no. March, p. 124391, 2024, doi: 10.1016/j.eswa.2024.124391.

[16] S. R. Dinesh Jackson *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Comput. Networks*, vol. 151, pp. 191–200, 2019, doi: 10.1016/j.comnet.2019.01.028.

[17] D. R. Patrikar and M. R. Parate, "Anomaly detection using edge computing in video surveillance system: review," *Int. J. Multimed. Inf. Retr.*, vol. 11, no. 2, pp. 85–110, 2022, doi: 10.1007/s13735-022-00227-8.

[18] A. Marwaha, A. Chirputkar, and P. Ashok, "Effective Surveillance using Computer Vision," *2nd Int. Conf. Sustain. Comput. Data Commun. Syst. ICSCDS 2023 - Proc.*, pp. 655–660, 2023, doi: 10.1109/ICSCDS56580.2023.10105124.

[19] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An efficient anomaly recognition framework using an attention residual lstm in surveillance videos," *Sensors*, vol. 21, no. 8, 2021, doi: 10.3390/s21082811.

[20] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks," *IEEE Access*, vol. 7, pp. 39172–39179, 2019, doi: 10.1109/ACCESS.2019.2906275.

[21] I. Mugunga, J. Dong, E. Rigall, S. Guo, A. H. Madessa, and H. S. Nawaz, "A frame-based feature model for violence detection from surveillance cameras using ConvLSTM network," *2021 6th Int. Conf. Image, Vis. Comput. ICIVC 2021*, pp. 55–60, 2021, doi: 10.1109/ICIVC52351.2021.9526948.

[22] S. M. Muiruri, M. Okong'o, and D. Mwathi, "Enhancing Public Safety Through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage," *East African J. Inf. Technol.*, vol. 7, no. 1, pp. 202–214, 2024, doi: 10.37284/eajit.7.1.2117.

[23] F. U. M. I. N. Ullah and S. Korea, "A Comprehensive Review on Vision-Based Violence," vol. 55, no. 10, 2023.

[24] H. Mohammadi and E. Nazerfard, "Video violence recognition and localization using a semi-supervised hard attention model," *Expert Syst. Appl.*, vol. 212, no. August 2022, p. 118791, 2023, doi: 10.1016/j.eswa.2022.118791.

[25] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies," *Appl. Sci.*, vol. 9, no. 22, 2019, doi: 10.3390/APP9224963.

[26] A. Pandey and P. Kumar, "Resstanet: deep residual spatio-temporal attention network for violent

action recognition," *Int. J. Inf. Technol.*, vol. 16, no. 5, pp. 2891–2900, 2024, doi: 10.1007/s41870-024-01799-w.

[27] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained Deep Ordinal Regression for End-to-End Video Anomaly Detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 12170–12179, 2020, doi: 10.1109/CVPR42600.2020.01219.

[28] S. A. Sumon, R. Goni, N. Bin Hashem, T. Shahria, and R. M. Rahman, "Violence Detection by Pretrained Modules with Different Deep Learning Approaches," *Vietnam J. Comput. Sci.*, vol. 7, no. 1, pp. 19–40, 2020, doi: 10.1142/S2196888820500013.

[29] S. Subramani, H. Wang, H. Q. Vu, and G. Li, "Domestic violence crisis identification from facebook posts based on deep learning," *IEEE Access*, vol. 6, pp. 54075–54085, 2018, doi: 10.1109/ACCESS.2018.2871446.

[30] C. L. MacIver *et al.*, "Macro- and micro-structural insights into primary dystonia: a UK Biobank study," *J. Neurol.*, vol. 271, no. 3, pp. 1416–1427, 2024, doi: 10.1007/s00415-023-12086-2.

[31] J. Kukad, S. Soner, and S. Pandya, "Autonomous Anomaly Detection System for Crime Monitoring and Alert Generation," *J. Autom. Mob. Robot. Intell. Syst.*, vol. 16, no. 1, pp. 62–71, 2022, doi: 10.14313/JAMRIS/1-2022/7.

[32] M. Hasanudin, A. H. Arribathi, Indrianto, K. Yuliana, and D. P. Kristiadi, "Increasing Independence of Cerebral Palsy Children using Virtual Reality based on Mlearning," *J. Phys. Conf. Ser.*, vol. 1764, no. 1, 2021, doi: 10.1088/1742-6596/1764/1/012119.

[33] D. Durães, B. Veloso, and P. Novais, "Violence Detection in Audio: Evaluating the Effectiveness of Deep Learning Models and Data Augmentation," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, no.3, pp. 72–84, 2023, doi: 10.9781/ijimai.2023.08.007.