NaturalResourcesfor
Human Health
eISSN: 2583-1194

Original Research

Received 12 March 2026
Revised 16 April 2026
Accepted 20 May 2026

Natural Resources for Human Health 2026, Vol. 6 (1s): 45–54

KEYWORDS:
LSTM, GRU, ANN, VGG16, MNAS, ROG-AUC

A Robust Deep Learning Approach for Real-Time Violence Detection: Comparative Study of CNN, LSTM, and GRU

Jyoti Kukade[1], Prashant Panse[2], Vidhya Barpha[3], Trapti Mishra[4], Rahul Singh Pawar[5]
[1,2,3,4,5] Medicaps University

ABSTRACT:
Time-to-real detection of potential threats has formed the base of maintaining public safety through violence detection systems. The present study dwells on deep learning models that offer improved prediction accuracy for human activity analysis with the changed scope of their AI use. Recent studies use a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs such as LSTM) in many settings to classify violence in videos obtained from surveillance footage, attaining a substantial drop in the rate of classification errors. This research constitutes a hybrid deep learning framework that combines CNNs for extracting features and LSTMs for analysing violent and non-violent scenes. It was also examined how well the GRUs identified and classified the scenes along with fully connected networks in parallel so as to increase the robustness of the model. CNNs extract spatial features, LSTMs and GRUs capture temporal dynamics. The experiment reaches 92% test accuracy, accomplishing an impressive advance in differentiating violent onset from benign events, achieving superior precision and recall values compared with state-of-the-art benchmarks. This finding delineates the contribution of deep learning in respect of the effective enhancement of automated surveillance, paving the path for safer environments through proactive presence of threat detection.

1. INTRODUCTION

Physical violence affects almost everyone's life, not only them but also their families and beloveds. It also destabilizes their economy and the mental health because of the insecurities caused by violence. There are some studies that come up with shocking findings, for instance, physical violence is the primary reason for death among individuals across the globe. The research and study conducted by Hillis et al.in 2016, which has been published in the American Academy of Pediatrics, indicates that over half of the children in continents such as Asia, Africa, and North America suffered violence assaults in the year 2015, and globally, over a half of all children under the age of 17 suffered such violence. Based on a research/study conducted by the European Union Agency for Fundamental Rights, 1 out of every 4 Europeans were victims of violence and approximately 22 million were physically attacked within 1 year. For such types of reasons, it's essential to resolve the problem of physical violence in societies.

There are long-term, medium-term, and short-term solutions that have been proposed. Long-term solutions include learning about causes or circumstances that lead to an escalation of violence, so that subsequent generations can be taught to avoid them. There have been some studies investigating an escalation of violence as a result of exposure to a violent and aggressive environment, and the use of violence as a means. There has been the illustration of aggression having a causative effect on an unhappy maternal relationship or low self-esteem. Additionally, technology, namely smartphones, is a reason for a rise in violence, as is the consumption of violent video games.

Other researchers have employed the medium-term solutions approach in addressing violence by attempting to link the density of the street population and the urban landscape to crime rates. The population density of the urban area and how it is linked to the crime rate have been measured by employing social network and mobile phone data and statistical methods geared toward spatial analysis, even though the weakness that it is not possible to distinguish the indoor and outdoor population is a limitation. Additionally, through the employment of convolutional neural networks (CNNs), crime rates have been seen to be reduced in green areas. Another work employed deep learning in evaluating images of Baidu Street View (Chinese version of Google Street View) of the street population number, the type of urban environment (buildings, green areas, etc.), and linked them to the number of crimes. Another study also used Google Street View images to analyze the number of vehicles and the pavement, and related those features to the crime rate through the use of machine learning techniques.

The subject of this review is that the short-lived strategies and rapid prevention of physical assault are the ones that operate in real time. As of yet, to provide post-crime evidence that is also used as an evidence to identify the offenders and used by insurance companies or police, images and security cameras have primarily been utilized. Although a considerable number of published papers are already available that have implemented real-time violence detection solutions.

These approaches are based on the use of security video cameras led by artificial intelligence (AI) algorithms. It is a new field whose development has been hastened by the advent of three pillars: the expansion in the utilization of images and security cameras, the technological advancement of platforms of big data, and the expansion of artificial intelligence algorithms that support the evaluation of images and video.

The use of security cameras has gone up across the world to be utilized for safety and monitoring, among others. It is critical to point out the fact that while the photos and videos possess different uses which can be harnessed for enhancing the living standards of the population, studies already exist calling into question the danger to individuals' privacy in mass recording.

Today, big data platforms provide us with the means to acquire and process large-scale data gathered from the world around us. These devices allow us to capture, store, and process, in real-time, different types of data, like images and videos. Lastly, image and video analysis-based artificial intelligence algorithms are now more common in recent years; they have also become more diverse and accurate.

Recognition of violent activities in videos falls under the category of computer vision, which is action recognition. Computer vision is a branch of artificial intelligence that allows computers to interpret and make decisions from visual data. Action recognition is a research area that deals with the recognition of specific activities within streams of video. AI-supported violence detection in videos is the training of models to recognize patterns and activities that are characteristic of violence.

AI-based violence detection from video is based on machine learning models trained from annotated video datasets where violent and non-violent acts are labeled and tagged. The algorithms are trained to identify patterns and features related to violent acts and therefore tag the same in unseen, new videos.

Overall, physical aggression is a significant social issue with widespread impact. There have been numerous research programs created to solve the problem in a variety of manners, and each has been providing solutions of varying approach scales. Real-time detection of violence acts is the fastest and most critical solution, and it is the ultimate means of protection in recognizing victims of physical harm. Its success is credited to a series of core factors: growing reliance on video and security footage, the enhancement of big data platforms and algorithms in a bid to process images and video with the support of artificial intelligence.

The goal of this research is to further develop a systematic mapping study with the aim of presenting a complete and timely overview of AI-powered video violence detection, targeting physical assault as a primary focus. The key contribution of this work lies in the formation of an extensive and recent review, addressing all the stages involved in video violence detection using artificial intelligence. Lastly, a factor that differentiates this paper is its evaluation of different algorithm types used in video violence detection, the set of algorithms employed together, and the outcomes achieved for the most frequently used datasets in the state-of-the-art literature.

2. LITERATURE REVIEW

Various Literature works are taken under consideration and studied in order to create an algorithm that is effective and robust for violence detection under various conditions. These works of literature are some that try to achieve similar goals and try different algorithms and models that are available and add some new approaches on top of that. They are as follows:

Talha et al. [8] proposed a highly effective and fast real-time violence detection system tested on the personal machines of its authors. The system relies on a CNN for extracting spatial functions, which may be fed into an LSTM. besides, the CNN has fully connected layers for classification functions. Madhavan [1] presented a variant to counter difficult conditions related to class in various weather and light fixture conditions, though overall performance metrics have not been provided. The variant handled issues like allotting a minimum number of pixels for video categories. Ullah et al. [2] employed Mask R-CNN, an extension of Faster R-CNN, to select keyframes by object detection and object labeling of humans and motors in images. DarkNet and any other CNN that employed optical float as residual input were used by them for function extraction. These features were examined with a multilayer long short-term memory (M-LSTM) network.

Vijeikis et al. [3] developed a light and fast model with CNN and LSTM, albeit with slightly less accuracy. Halder and Chatterjee [] employed a light-weight CNN-based bidirectional LSTM with satisfactory performance for violent activity recognition, with impressive results. Traoré and Akhloufi [4] employed a pre-trained VGG-sixteen with the INRA character database for spatial characteristic feature extraction. those features were directly fed to a bidirectional gated recurrent unit (BiGRU), along with employing 3 connected layers, with the final layer employing softmax activation. In addition, Ref. [4] employed VGG-16 for spatial feature extraction and LSTM for temporal feature analysis.

Asad et al. [5] proposed a violence detection model that processed consecutive video frames at times t and t+1 as inputs, rather than their difference. Two pre-trained CNNs were used for high- and low-level feature extraction from these frames. extensive dense residual blocks (WDRBs) then blended these features, and an LSTM captured temporal styles among them. The gadget also plotted an actual-time graph to signify violence tiers, raising an alarm while the output handed a threshold. Contardo et al. [6] utilized MobileNetV2, a CNN pre-trained on ImageNet, to extract spatial features. These outputs had been fed into two kinds of LSTMs—temporal Bi-LSTM and temporal ConvLSTM—to evaluate their relative performance.

Jahlan and Elrefaei et al. [7] used random body selection per packet and recompressed the images into squares by cropping different portions. Rather than capturing character frames, they processed the difference between two consecutive frames (i and that i+1). They followed MNAS (cellular Neural architecture seek), a lightweight CNN, paired with convolutional LSTM for temporal and spatial characteristic extraction. The extracted capabilities were normalized between zero and 1 earlier than being categorized by the usage of 3 exclusive classifiers, with SVM turning in high-quality outcomes. Islam et al. [9] focused on detecting sexual and physical assaults while providing designated parameters of the datasets, which includes elegance remember, video depend, frames in step with 2d, video duration, average frames per video, decision, and vicinity statistics. Their model leveraged pre-educated VGG-sixteen and VGG-19 networks, with outputs fed into LSTMs.

Thanjaivadivel T, Dinesh Jackson Samuel R, Fenil E, Gunasekaran Manogaran, Vivekananda G.N, Ahilan A, and Jeeva S[10] developed a concurrent detection framework for violence in January 2019 that is capable of processing massive video streams based on the Spark framework. It extracts frames and processes them using the Histogram of Oriented Gradients (HOG) function, classifies them based on a violence model, human part model, and negative model, and trains a Bidirectional Long Short-Term Memory (BD-LSTM) network. Using bidirectional information flow, BD-LSTM generates outputs based on previous and subsequent context. It was trained using a dataset that comprised 2,314 videos with 1,077 fight and 1,237 non-fight scenes, and a specially designed dataset with 410 non-violent and 409 violent video clips extracted from football stadium videos.

Srivastava et al. [11] suggested two algorithms, which are described below. The primary was the violence classification model. The secondary was a facial verification system beneficial for violence detection. This algorithm emphasized the detection of violence integrating drone surveillance. Spatial patterns were identified by applying a CNN unit that used pre-built ImageNet-based architectures, whose result is input into the LSTM. Overall, seven distinct algorithms were utilized, in addition to the merged version of some algorithms. Akhloufi and Traoré[12] have employed two CNNs, named EfficientNet, pre-trained on ImageNet. One of these processes optical flow, while the other handles RGB. The outputs from both CNNs were fed into an LSTM, followed by a classifier featuring a fully connected layer(FCL) with a sigmoid activation function.

Ji et al. [13] proposed the Human Violence Dataset- a new dataset with 1930 downloaded movie trailer clips from YouTube. It was not restricted to physical aggression only but also to gun violence. Dual-stream CNN models are a specific neural network structure based on two distinct streams of visual information: spatial and temporal streams. A single image is taken as an input to a temporal stream and 10 optical flow frames are taken as an input to a spatial stream. Upon feature extraction via CNN, a computational process trains the classifier to increase the weights to evaluate the aggression level of the clips. This article quantified levels of violence in the videos in comparison with a ranking score. Therefore, three grades of aggression (L1, L2, and L3) were computed with the help of a confusion matrix.

Baba et al. [14] introduced a light-weight CNN model, i.e., SqueezeNet and MobileNet, on two public datasets, i.e., the BEHAVE and ARENA datasets. To improve the performance of the model, a time-domain filter was applied to distinguish violent scenes within a given time window. The overall experimental result achieved an accuracy rate of 77.9% with no misclassifications in the violent class but a very high rate of false positives equivalent to 26.76% of the non-fight clips. The authors also mentioned that one of the major limitations of the proposed method is that the proposed method fails to detect violence in crowded scenes.

CNN and LSTM were recently designed to discover violence in video features. CNN is used for keypoint extraction at frame level. The extracted features are categorized as either violent or non-violent by utilizing a variant of LSTM in the framework by Sudhakaran and Lanz[15], Soliman et al.[16], Letchmunanet al.[17], and Sumon et al.[18]. Through CNN and LSTM integration, the videotape's spatiotemporal features are localized, enabling precise motion analysis. Diverse pre-trained CNN models, VGG19 by Letchmunanet al.[17], videlicet VGG16 by Soliman et al. [16], and ResNet50 by Sumon et al.[18] were utilized to extract spatial features, before their classification as violent or non-violent.

Hanson et al.[19] presented a model divided into three-path spatial encoders, temporal encoders, and classifiers. The authors previously used ConvLTSM infrastructures reinforced with bidirectional temporal encodings. The model showcased is known as a one-sluice model, as it utilizes only one pattern for input; nonetheless, some workshops showcased models that use both formats for their input. These models are called convolutional multistream models, where the type of each videotape sluice is anatomized as well.

Shakil Ahmed Sumon, Tanzil Shahria, Niyaz Bin Hashem, Raihan Goni, and RashedurM[20] . Rahman discussed many methods for selection of prominent features from different pre-trained models in order to classify violence in a video. The authors created a dataset contrasting violent and non-violent video excerpts and used VGG16, VGG19, and ResNet50 in order to pick features from a frame. In one method, feature-extracted features were passed through a fully connected network for violence detection at the frame level, and in another experiment, features from 30 consecutive frames were processed using an LSTM network. They also employed attention mechanisms via spatial transformer networks to enhance feature extraction by allowing features such as rotation, translation, and scaling. In addition to these models, they proposed a bespoke CNN as a feature extractor and integrated a pre-trained model that was first trained on a movie violence dataset. Of all the methods, ResNet50 features performed the best, and when coupled with an LSTM network, provided a better accuracy of 97.06%, surpassing other models that were tested.

Rohit Halder and Rajdeep Chatterjee et al. [21] introduced a featherlight computational model for the improved classification of violent and non-violent behavior. Deep literacy grounded in an effective violent exertion discovery model can assist authorities in monitoring violent activity in real-time. The assessed results can be hereafter moved to store and examine the footage to streamline the crime surveillance system. CNN-based Bidirectional LSTM has been utilized to explain violent behavior adaptation and additionally contrasted with other methods.

Hua et al.[22] proposed improving human pose estimation by a residual attention mechanism added to layered hourglass architecture for resolution loss in initial images. Their architecture improved resolution and accuracy, which resulted in more precise posture detection in images with dense background. Liu et al.proposed a novel method of preserving temporal coherence of human posture detection between video frames. Their approach utilized structured spatial learning and partial temporal analysis in three-phase multi-feature deep CNN for stable and precise long-term posture detection.

Magdy et al.[23] segmented video data into frame packets and used an optical flow-based method for motion area detection. Their research compared CNN-3D and CNN-4D architectures and concluded that CNN-3D efficiently processed short time periods, whereas CNN-4D offered a more sophisticated perspective of intricate space-time correlations. The CNN models were pre-trained on ImageNet first.

Sernaniet al.[24] emphasized the significance of reliability in hostility identification and introduced the AIRTLab dataset, purposefully created to assess algorithm resilience. Their study presented three different deep learning architectures, where pre-trained 3D-DL models, fine-tuned using the Sports-1M dataset, proved that feature reuse learning improves productivity over training models from the ground up.

Queet al.[25] worked on detecting violence in long videos using already-trained Deep Learning Models. Their study emphasized accurately determining the start and end of violent episodes, an area often overlooked in prior research. The second phase of their approach employed a deconvolution technique to pinpoint specific video segments at the frame level, precisely identifying when violent events occurred. The study also highlighted the need for improved preprocessing and detection methods for marking the beginning and end of violent activities.

Kaur and Singh et al.[26] epitomized contemporary reviews on violence discovery since 2016. They bifurcated the reviewed documents into the two main groups: traditional styles, further subdivided by point birth and bracket ways; and deep literacy styles. This research paper focused on studies using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) as current and promising ways in this field. Colorful challenges in violence discovery were bandied, comprising issues similar to changes in illumination, and lapping objects. The paper deduced that there was no outcome to these challenges inclusively.

Yao and Hu et al.[27] raised the defiance of characterizing viciousness because of its fundamental nebulosity and incidental circumstance, making genuine case recordings difficult to pick up. They bandied about the contrast between ordinary styles and profound proficiency (DL) ways for recognizing hostility but didn't allow points of interest almost the composition choice handle, and counting databases. The audit scattered workmanship papers into two segments: the conventional outline and profound education. Applying CNN authors have classified the events into anomalous or non-anomalous events. They take videos as input process them in frames and perform this binary classification. Although, they were unable to identify outliers in it[30].

The design of the deep learning model was proposed to video anomaly detection. Here author used extracting spatial features and LSTM for temporal dependencies between the frames. Significant accuracy was achieved on the UCSD pedestrian dataset[31].

A comparative framework for predicting earthquake magnitude was created using machine learning algorithms. This includes XGBoost, Stacking Regressor, Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), LSTM, and Random Forest. Historical seismic data were used to train and assess the models. The process included optimizing parameters and evaluating performance with standard regression metrics. The study focused on a data-driven approach for analyzing seismic hazards and predicting earthquakes. The author proposed a deep learning approach for detecting crime through automatic driving vehicles. They have used OpenCV for object detection and LSTM for classification which results in surveillance of live streams like mob lynching or burglary. They have achieved 90% accuracy through their system [32].

[33] To enhance security, transparency, and verifiability in digital elections, a blockchain-based e-voting framework was built. With an integrated architecture of Ethereum smart contracts and Multichain blockchains, this system advanced the decentralized recording of votes and the automatic counting of votes. Secure voter authentication was established through decentralized identity management, and integrity and auditability of votes were guaranteed through immutable blockchain storage. The proposed architecture proved the capability of blockchain technology in designing scalable and reliable electronic voting systems [34].

3. METHODOLOGY:

This work combines LSTM, GRU, and ANN structures with the InceptionV3 pre-trained network to create an efficient and accurate system for violence detection in video streams. The research uses the Real Life Violence Dataset, consisting of video segments labeled as Violence and Non-Violence. In preprocessing, a maximum of 12 representative frames per video are extracted, resized to 299×299 pixels (the requirement for InceptionV3), and pixel values are normalized to the range of 0 and 1 to preserve uniformity. To obtain significant spatial features from the video frames, the InceptionV3 model, pre-trained on ImageNet, is used. The fully connected layers of InceptionV3 are removed, enabling feature representations to be obtained from the last convolutional layer, thus providing 2048-dimensional feature vectors for each frame.

To integrate temporal dependencies into the classification model, the feature vectors derived are fed into LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) models, both of which are effective in modeling sequential data. The

LSTM model takes the input of 12 frames × 2048 features and uses dropout regularization (0.2) as an overfitting countermeasure. Likewise, the GRU model uses a similar setup but provides a computationally more efficient alternative to LSTM, which is effective in capturing long-term dependencies with low complexity. Alternatively, the ANN (Artificial Neural Network) model uses a non-recurrent setup by flattening the extracted features and feeding them into fully connected dense layers. The ANN model uses several layers, including Dense(512, ReLU) → Dropout(0.2) → Dense(256, ReLU) → Dropout(0.2) → Dense(128, ReLU) → Dropout(0.2), producing a softmax activation output layer for binary classification.

The models are optimized using the Adam optimizer with a learning rate of 0.001, along with the use of a sparse categorical cross-entropy loss function. The models are trained for 20 epochs with a batch size of 30 to ensure optimal convergence. Model performance is evaluated using key metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curves. After training, a comparative analysis between LSTM, GRU, and ANN models is performed to determine the best architecture for real-time violence detection. The last trained model is saved and deployed, thereby allowing the automation of violence detection in video feeds. The use of deep feature extraction (InceptionV3) combined with sequential modeling (LSTM, GRU) and fully connected networks (ANN) ensures a robust and accurate classification system, with computational efficiency and high performance.

Apart from the early models, the current study further explores the incorporation of VGG19, InceptionV3, and NASNetMobile with Artificial Neural Networks (ANN) to enhance the detection of violence in video content. All these pre-trained convolutional neural networks (CNNs) were utilized for feature extraction, followed by classification using a fully connected ANN. The VGG19 model, renowned for its deep architecture with 19 layers, was utilized for extracting hierarchical spatial features, whereas InceptionV3, renowned for its effective multi-scale convolutions, was utilized to extract more abstract feature representations. In addition, NASNetMobile, an AutoML-based CNN, was utilized to take advantage of its optimized framework for lightweight and efficient feature extraction.

The feature vectors of the final convolutional layers of all these CNNs were flattened and passed through an ANN classifier, which comprised a number of dense layers with ReLU activation functions and dropout regularization to prevent overfitting. The ANN architecture remained the same in all three models to provide a fair comparison of performance. The training process was the same as before with a same optimization strategy, which used the Adam optimizer with a learning rate of 0.001 and sparse categorical cross-entropy loss with 20 epochs. Performance evaluation was done with accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC curves to quantify the performance of every one of the feature extraction techniques. The results were compared to the current LSTM, GRU, and ANN models, and insights were gained on the trade-offs involved with sequential modeling (LSTM/GRU) versus solely ANN-based techniques with varying CNN backbones. The top-performing model was ultimately used to develop an efficient and accurate real-time violence detection system.

Hyperparameter tuning was implemented on these three models to achieve better accuracy as compared to the prior activation functions like Softmax. We have used Relu, Selu, Elu, Sigmoid, Tanh, Adam as activation functions for the hyperparameter tuning process.

4.RESULTS AND DISCUSSION

4.1 Performance evaluation on Fight Detection Dataset

In order to test the performance of the suggested fight detection system, three deep learning models were employed, which are Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) Network, and Gated Recurrent Unit (GRU). The assessment utilized accuracy and loss metrics to identify the appropriate model to use in detecting fights in video surveillance in real-time.

Figure 1 shows the CNN model's accuracy and loss, which were 86.28% and 36.56, respectively. The fluctuation of accuracy and loss over epochs shows a consistent trend of convergence, which indicates that CNN can learn patterns related to fight from video sequences well.

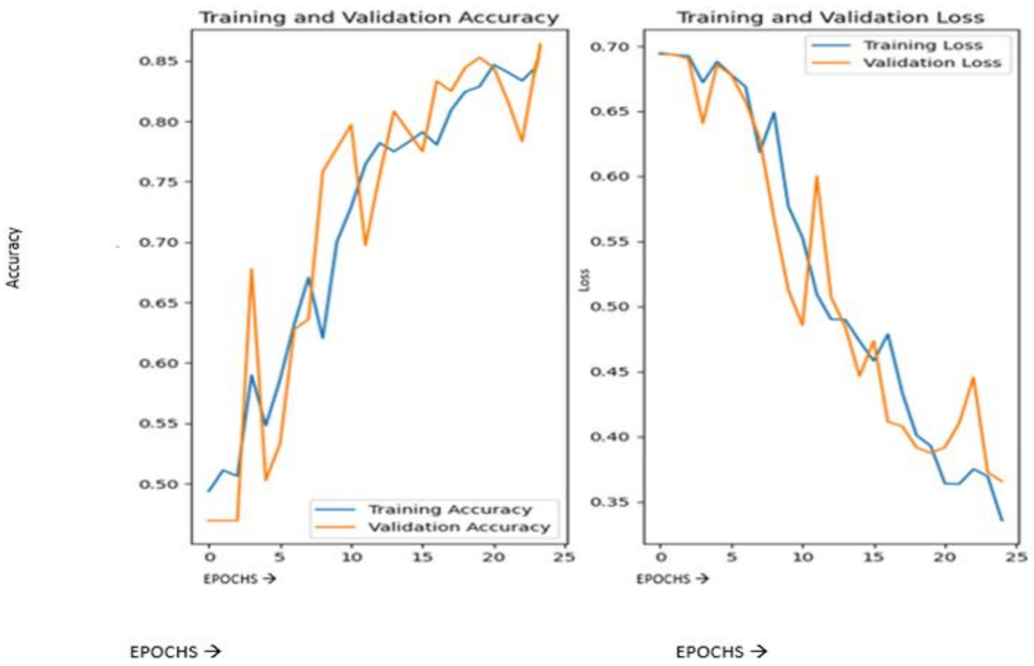

Figure 1 Performance of CNN for Fight Detection System (a) Accuracy (b) Loss variations with epochs

Likewise, Figure 2 shows the efficiency of the LSTM model. The LSTM was 78.29% accurate and had a loss of 52.11. Although accuracy improved consistently over epochs, the loss curve is steeper, indicating issues in learning long-term dependency for sequential fight detection.

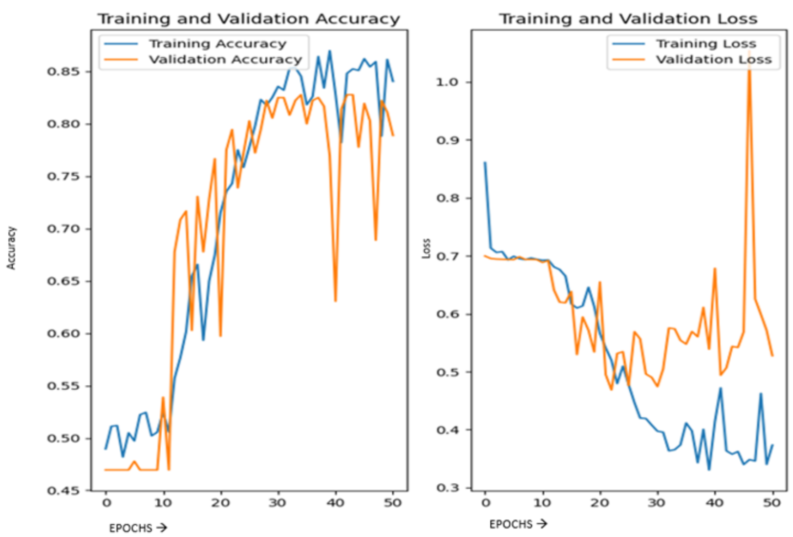

Figure 2 Performance of LSTM for Fight Detection System (a) Accuracy (b) Loss variations with epochs

Figure 3 is a demonstration of GRU's performance, as it is the best of the three models with an accuracy of 91.01% and a smaller loss factor of 34.3. The stability of the accuracy curve and the decreasing loss trend signify GRU's effectiveness in capturing temporal dependencies while maintaining lower computational complexity.

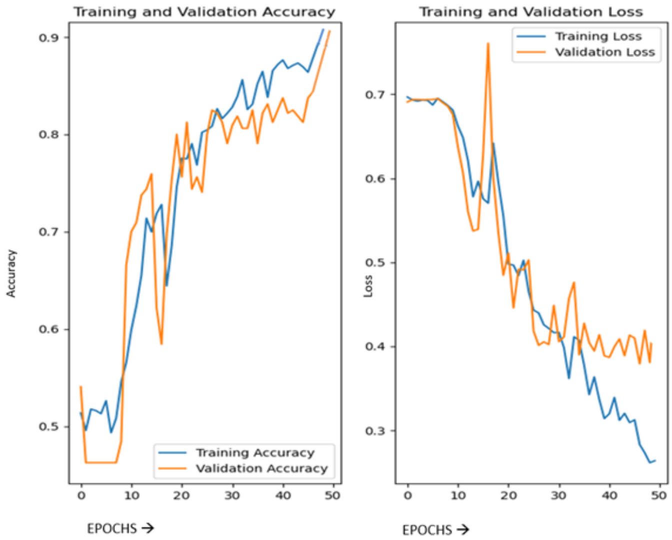

Figure 3 Performance of GRU for Fight Detection System (a) Accuracy (b) Loss variations with epochs.

4.2 Comparative Performance Analysis

In conclusion, the study discovers GRU as the highest performing model in fight detection. The use of CNN, LSTM, and GRU models results in an improved system of violence detection for video surveillance systems. The result confirms that GRU performs better than other techniques in the literature and is a viable choice for real-time use in automated crime surveillance systems.

This research verifies that deep learning models, especially GRU, greatly improve the accuracy of violence detection. Future research can be directed towards model generalization to other video datasets and real-time deployment effectiveness.

5. CONCLUSION

This work contributes to real-time violence detection by building a hybrid deep learning model incorporating CNNs, LSTMs, and GRUs. The model attains 92% test accuracy, an improvement over existing state-of-the-art. Real-time deployment optimizations and privacy-preserving methods will be investigated in future work to promote ethical AI surveillance.

Its future potential for the detection of violence includes improving accuracy and operational performance in real-time detection systems using the implementation of novel deep learning architectures and edge AI. Multi-modal inputs, e.g., speech and biometric, can be utilized to further enhance classification performance. Its use in public surveillance, smart city infrastructure, and law enforcement can facilitate proactive deterrence of crime. Further model optimization to reduce false positive rates and its deployment in diverse environments such as schools and workplaces, would further improve its real-world utility.

ACKNOWLEDGEMENTS

The authors sincerely acknowledge Medicaps University, Indore, and the Department of Information Technology for providing the necessary academic support, infrastructure, and resources to carry out this research work. The authors also extend their gratitude to the faculty members and colleagues for their valuable guidance, suggestions, and encouragement throughout the research.

AUTHOR CONTRIBUTIONS

Jyoti Kukade contributed to conceptualization, methodology development, data collection, experimentation, and preparation of the original manuscript. Prashant Panse contributed to the literature review, data preprocessing, investigation, validation, and interpretation of the experimental results. Vidhya Barpha was involved in software implementation, formal analysis, visualization, performance evaluation, and documentation of the results. Trapti Mishra contributed to research supervision, resource management, result verification, and critical review of the manuscript. Rahul Singh Pawar contributed to conceptualization, research methodology, technical supervision, project administration, interpretation of findings, and critical revision of the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This research follows accepted ethical standards. All data were obtained from publicly available or authorized sources and used solely for academic research purposes.

DATA AVAILABILITY STATEMENT

Not Applicable

REFERENCES

[1] Madhavan, R. & Utkarsh, & Vidhya, J.. (2021). Violence Detection from CCTV Footage Using Optical Flow and Deep Learning in Inconsistent Weather and Lighting Conditions. 10.1007/978-3-030-81462-5_56.

[2] Ullah, Amin & Muhammad, Khan & Del Ser, Javier & Baik, Sung & Albuquerque, Victor. (2018). Activity Recognition using Temporal Optical Flow Convolutional Features and Multi-Layer LSTM. IEEE Transactions on Industrial Electronics. PP. 1-1. 10.1109/TIE.2018.2881943.

[3] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," Sensors, vol. 22, no. 6, p. 2216, Mar. 2022. doi: 10.3390/s22062216.

[4] L. A. Siddique, R. Junhai, T. Reza, S. S. Khan, and T. Rahman, "Analysis of Real-Time Hostile Activity Detection from Spatiotemporal Features Using Time Distributed Deep CNNs, RNNs and Attention-Based Mechanisms," arXiv preprint arXiv:2302.11027, 2023. [Online]. Available:https://arxiv.org/abs/2302.11027.

[5] Asad, Mujtaba & Yang, Jie & He, Jiang &Shamsolmoali, Pourya & He, Xiangjian. (2021). Multi-frame feature-fusion-based model for violence detection. The Visual Computer. 37. 10.1007/s00371-020-01878-6.

[6] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo and A. F. Dragoni, "Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset," in IEEE Access, vol. 9, pp. 160580-160595, 2021, doi: 10.1109/ACCESS.2021.3131315.

[7] H. M. B. Jahlan and L. A. Elrefaei, "Detecting violence in video based on deep features fusion technique," arXiv preprint, arXiv:2204.07443, 2022. [Online]. Available:https://arxiv.org/abs/2204.07443

[8] K. R. Talha, K. Bandapadya, and M. M. Khan, "Violence detection using computer vision approaches," in 2022 IEEE World AI IoT Congress (AIIoT), IEEE, 2022, pp. 544–550.

[9] P. Negre et al., "Literature Review of Deep-Learning-Based Detection of Violence in Video," Sensors (Basel, Switzerland), vol. 24, no. 12, p. 4016, Jun. 2024. doi: 10.3390/s24124016.

[10] Samuel, Dinesh Jackson & Fenil, E. &Manogaran, Gunasekaran &Gn, Vivekananda &Thanjaivadivel, M. & Selvaraj, Jeeva &Appathurai, Ahilan. (2019). Real time Violence Detection Framework for Football Stadium comprising Big Data Analysis and Deep Learning through Bidirectional LSTM. Computer Networks. 151. 10.1016/j.comnet.2019.01.028.

[11] Srivastava, Anugrah & Badal, Tapas & Singh, Rishav. (2021). Real Life Violence Detection in Surveillance Videos using Spatiotemporal Features. 262-266. 10.1145/3474124.3474161.

[12] A. Traoré and M. A. Akhloufi, "2D Bidirectional Gated Recurrent Unit Convolutional Neural Networks for End-to-End Violence Detection in Videos," in Proceedings of the International Conference on Artificial Intelligence Applications and Innovations (AIAI), 2020. doi: 10.1007/978-3-030-50347-5_14.

[13] Negre, Pablo et al. "Literature Review of Deep-Learning-Based Detection of Violence in Video." Sensors (Basel, Switzerland) vol. 24,12 4016. 20 Jun. 2024, doi:10.3390/s24124016

[14] Baba M., Gui V., Cernazanu C., Pescaru D. A sensor network approach for violence detection in smart cities using deep learning. Sensors. 2019;19:1676. doi: 10.3390/s19071676.

[15] Sudhakaran, Swathikiran& Lanz, Oswald. (2017). Learning to detect violent videos using convolutional long short-term memory. 1-6. 10.1109/AVSS.2017.8078468.

[16] M. Soliman, M. A. Ismail, and M. A. El-dosuky, "Robust Real-Time Violence Detection in Video Using CNN And LSTM," arXiv preprint arXiv:2107.07578, 2021. [Online].

[17] S. Letchmunan, S. K. Subramaniam, and M. K. Lim, "A Fully Integrated Violence Detection System Using CNN and LSTM," International Journal of Electrical and Computer Engineering (IJECE), vol. 11, no. 4, pp. 3374-3380, 2021.

[18] S.I.Sumon, "Human Activity Recognition using CNN and LSTM," GitHub Repository,2020.[Online].Available:https://github.com/shafiqul-islam-sumon/HARCNNLSTM

[19] Hanson, A., PNVR, K., Krishnagopal, S., Davis, L.: Bidirectional Convolutional LSTM for the Detection of Violence in Videos. Lecture Notes in Computer Science 2019, 280-295.

[20] S. A. Sumon, R. Goni, N. B. Hashem, T. Shahria, and R. M. Rahman, "Violence Detection by Pretrained Modules with Different Deep Learning Approaches," Vietnam Journal of Computer Science, vol. 7, no. 1, pp. 19-40, 2020, doi:10.1142/S2196888820500013.

[21] R. Halder and R. Chatterjee, "CNN-BiLSTM Model for Violence Detection in Smart Surveillance," SN Computer Science, vol. 1, 2020. Available:https://api.semanticscholar.org/CorpusID:220794917.

[22] Gao, Y., Liu, H., Sun, X., Wang, C. and Liu, Y.: Violence detection using Oriented Violent Flows. Image and Vision Computing, vol. 48-49, 37-41 (2016).

[23] Magdy, M., Fakhr, M.W., Maghraby, F.A.: Violence 4D: violence detection in surveillance using 4D convolutional neural networks. IET Comput. Vis. 17(3), 282–294 (2023). https://doi.org/10.1049/cvi2.12162

[24] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo and A. F. Dragoni, "Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset," in IEEE Access, vol. 9, pp. 160580-160595, 2021, doi: 10.1109/ACCESS.2021.3131315.

[25] Qu, W.; Zhu, T.; Liu, J.; Li, J. A time sequence location method of long video violence based on an improved C3D network. J. Supercomput. 2022, 78, 19545–19565.

[26] Kaur, Harjinder et al. "A Review of Machine Learning based Anomaly Detection Techniques." ArXiv abs/1307.7286 (2013)

[27] Yao, Huiling and Xing Hu. "A survey of video violence detection." Cyber-Physical Systems 9 (2021): 1 - 24.

[28] H. Hillis, "Global Prevalence of Past-year Violence Against Children: A Systematic Review and Minimum Estimates," American Academy of Pediatrics, vol. 137, no. 3, pp. e20154079, 2016. doi: 10.1542/peds.2015-4079.

[29] European Union Agency for Fundamental Rights, Violence Against Women: An EU-Wide Survey – Main Results Report, 2014. [Online]. Available: https://fra.europa.eu/en/publication/2014/violence-against-women-eu-wide-survey-m ain-results-report.

[30] Kukade, Jyoti, and Prashant Panse. "Advanced Deep Learning Model for Anomaly Detection Based Video Surveillance System." International Journal of Intelligent Systems and Applications in Engineering 12, no. 5s (2024): 477-485.

[31] Kukade, Jyoti, and Prashant Panse. "Designing a Deep Learning Model for Video Anomaly Detection-Based Surveillance." In International Conference on ICT for Sustainable Development, pp. 257-269. Singapore: Springer Nature Singapore, 2023.

[32] Kukade, Jyoti, Swapnil Soner, and Sagar Pandya. "Autonomous anomaly detection system for crime monitoring and alert generation." Journal of Automation, Mobile Robotics and Intelligent Systems 16, no. 1 (2022): 62-71.

[33] R. S. Pawar, P. Panse, J. Kukade, S. Neema, G. Singh and H. Goyal, "Leveraging Machine Learning Models for Seismic Prediction: A Study on Model Performance and Optimization," 2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON), Pune, India, 2025, pp. 01-08, doi: 10.1109/I2ITCON65200.2025.11210475.

[34] R. S. Pawar, P. Panse, J. Kukade, V. S. Barpha and A. Saxena, "Enhancing Transparency and Security in Elections: A Blockchain-Based E-Voting System(eVoteChain)," 2025 International Conference on Engineering Innovations and Technologies (ICoEIT), Bhopal, India, 2025, pp. 1562-1567, doi: 10.1109/ICoEIT63558.2025.11211646.