Violence Detection in Video Streams Using LSTM, GRU, and ANN with Deep Feature Extraction

Raj Narayan Singh Chouhan
Department of Information Technology
Medi-Caps University
Indore, India
0009-0000-2018-7333

Piyush Motwani
Department of Information Technology
Medi-Caps University
Indore, India
0009-0000-1107-1160

Raghav Mehrotra
Department of Information Technology
Medi-Caps University
Indore, India
0009-0009-5855-9687

Yuvraj Kocher
Department of Information Technology
Medi-Caps University
Indore, India
0009-0009-6312-8910

Prashant Panse
Department of Information Technology
Medi-Caps University
Indore, India
0000-0001-8129-1214

Jyoti Kukade
Department of Information Technology
Medi-Caps University
Indore, India
0009-4630009-3-0965

***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 fusion of convolutional neural networks (CNNs) and recurrent neural networks (Long Short Term Memory(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 analyzing 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 analyzed with the latest benchmarks. This finding delineates the influence of deep learning in terms of significant enhancement of automated surveillance, paving the path for safer environments through the proactive presence of threat detection.

***Index Terms*—LSTM, GRU, ANN, VGG16, MNAS, ROG-AUC**

I. 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.[27] in 2016, reported in the American Academy of Pediatrics indicates that over half of the children in continents such as Africa, North America, and Asia underwent brutal attacks in the year 2015, and globally, over a half of all children under the age of 17 suffered such violence. According to a study by the European Union Agency for Fundamental Rights[28], 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. Various solutions include short-term, medium-term, and long-term approaches. Long-term solutions include learning about causes or conditions that trigger an increase in violence, enabling future generations to learn how to avoid them. Research has explored how exposure to a violent and aggressive environment contributes to the use of violence as a means. Aggression has been discovered to have a causative role to perform in tense mother-child relationships and low self-esteem. Technology, particularly smartphones, has also been discovered to be a causative agent in promoting violence, as well as the playing of violent video games. Other researchers have employed the medium-term solutions approach in addressing violence by attempting to link the impact of urban density and landscape on 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 techniques specifically tailored for spatial analysis have been utilized, with the main limitation being the inability to distinguish between outdoor and indoor populations. Moreover, the use of convolutional neural networks (CNNs) has shown crime rates to decrease in green spaces. Another study used deep learning to investigate street-level imagery from Baidu, measuring street population density and urban environment attributes (e.g., buildings and parks) and relating them to crime rates. Another study also used Google Street View images to investigate vehicle numbers and pavement conditions using machine learning methods to determine their relationship with crime rates. The subject of this review is that the short-lived strategies and rapid control 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 methods are based on artificial intelligence (AI) powered security video cameras. This new technology has developed at a fast pace because of three major drivers: the high use of security cameras and image-based surveillance, improvements in big data platforms, and progressive developments in AI algorithms for visual data processing. Security cameras have found extensive use globally, and their purposes range from safety to surveillance. It is important to highlight 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. Contemporary big data platforms support capturing and processing voluminous data generated from our environments. Such frameworks support real-time capture, storage, and computation of multiple forms of data such as images and videos. There has been more widespread use in recent years of AI image and video analysis algorithms, with much diversity and a higher degree of accuracy.

Identifying violent activities in videos is a subset of computer vision, specifically action recognition. As a branch of AI, computer vision allows machines to interpret and make decisions based on visual information. Motion detection is a research area which deals with the recognition of specific activities within streams of video. Video violence detection using AI entails the process of training models to recognize patterns and movements characteristic of violent conduct. Machine learning violence detection from videos employs annotated datasets where violent and non-violent activities are labeled for training.. 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 violent 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, and the enhancement of big data platforms and algorithms in a bid to process images and video with the support of AI. The objective of this research is to move forward with a systematic mapping study to offer an extensive and latest overview of video violence detection by AI, the primary focus of which will be physical assault. The key contribution of this paper is creating an extensive and recent review that encapsulates all the phases of video violence detection using AI. 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.

II. Literature Review

Various Literature works are taken under consideration and studied 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:

A fast and effective real-time violence detection system was proposed by Talha et al. [8] and evaluated on the personal computers of its authors. The model 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 dealt with issues like distributing a minimum amount of pixels to various video classes. Ullah et al. [2] used Mask R-CNN, which is an enhanced version of R-CNN, for keyframe selection by identifying and annotating objects, i.e., humans and cars, from images. They utilized DarkNet along with other CNN models that leveraged optical flow as residual input for feature extraction. The extracted features were subsequently processed using a multilayer long short-term memory (M-LSTM) network.

Vijeikis et al. [3] built a light and fast model with LSTM and CNN, albeit with slightly less accuracy. Halder and Chatterjee [21] employed a light-weight CNN-based bidirectional LSTM with satisfactory performance for violent activity recognition, with impressive outcomes. Traoré and Akhloufi [4] employed a pre-trained VGG-16 combined with an INRA character database for extraction of spatial features.The features obtained from the extraction were fed directly into a bidirectional gated recurrent unit (BiGRU) with three coupled layers, of which the last was activated by softmax. Ref. [4] employed VGG-16 for spatial feature extraction as well as LSTM for temporal feature analysis.

Asad et al. [5] presented a violence detection method that utilized consecutive video frames at time t and t+1 as inputs instead of computing their differences. Two pre-trained CNNs were employed to extract high- and low-level features 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. MobileNetV2, a CNN pre-trained on ImageNet, was utilized by Contardo et al. [6] 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 archi-

tecture 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. [17] 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-16 and VGG-19 networks, with outputs fed into LSTMs.

Samuel, Dinesh Jackson & Fenil, E. and Manogaran, Gunasekaran & Gn, Vivekananda & Thanjaivadivel, M. and Selvaraj, Jeeva & Appathurai, Ahilan[9] developed a parallel 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 by applying the Histogram of Oriented Gradients (HOG) technique, categorizes them according to a violence model, human part model, and negative model, followed by training 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. [10] 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 by integrating drone surveillance. Spatial patterns were identified by applying a CNN unit that used prebuilt 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[11] have employed two CNNs, named EfficientNet, trained in advance on ImageNet. One of these handles optical flow processing, while the other handles RGB. The outputs from both CNNs were fed into an LSTM, followed by a fully connected layer (FCL) -based classifier with a sigmoid activation function.

Ji et al. [12] presented the Human Violence Dataset that contains 1,930 movie trailer clips derived from YouTube. The dataset includes instances of both physical aggression and gun violence. Dual-stream CNN models have a dedicated neural network architecture featuring two distinct streams of visual processing: spatial and temporal. 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. [13] introduced a light-weight CNN model, i.e., SqueezeNet and MobileNet, on two public datasets. A time-domain filter was used to improve model performance by recognizing violent scenes in a given time window. Experimental results indicated 77.9% accuracy without any misclassifications in the violent category. The model exhibited a high false positive rate, incorrectly classifying 26.76% of non-fight clips as fight clips. The authors also noted an important limitation: the suggested method has difficulty detecting violence in crowded scenes.

Recently, CNN and LSTM have been utilized for video feature violence detection. CNN is utilized for keypoint extraction in frames, while the extracted keypoint features are detected as violent or non-violent through the utilization of an LSTM variant in approaches introduced by Sudhakaran and Lanz [14], Soliman et al. [15], Letchmunan et al. [16], and Sumon et al. [17]. With the combination of CNN and LSTM, spatiotemporal features of a video are successfully localized, making accurate motion analysis easier. Hanson et al. [18] suggested a three-component model: spatial encoders, temporal encoders, and classifiers. The authors previously used ConvLSTM architectures augmented with bidirectional temporal encodings. The model introduced in this paper is called a one-sluice model and accepts a single input pattern. There are some works that have introduced models that handle multiple input forms, called convolutional multistream models, in which each type of video stream is processed individually.

Tanzil Shahria, Niyaz Bin Hashem, Shakil Ahmed Sumon, Raihan Goni, and Rashedur M[19]. They 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 VGG19, ResNet50, and VGG16 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 97.06Rohit Halder and Rajdeep Chatterjee et al. [20] proposed an efficient computational model to improve the classification of violent and non-violent actions. Deep literacy grounded in an effective violent exertion discovery model can assist authorities in live surveillance violent incidents. The tested results can then be stored and processed to improve the efficiency of the crime surveillance system. A CNN-based Bidirectional LSTM has been utilized to simulate violent behavior adaptation and has also been compared with other methods.

Hua et al.[21] 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 imaginary posture detection in dense backgrounds. Liu et al.proposed a method of preserving the profane 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.[22] segmented video data into frame sequences and used a velocity field method for motion area detection. Their research differentiated 3 3-dimensional CNN and 4-dimensional CNN architectures and concluded that 3-dimensional CNN efficiently processed short time periods, whereas 4-dimensional CNN offered a more sophisticated perspective of intricate space-time correlations. The CNN models were pre-trained on ImageNet first.

Sernani et al.[23] 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.

Que et al.[24] 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.[25] epitomized contemporary reviews on violence discovery since 2016. They bifurcated the documents into two batches: conventional styles, further branched by point birth and bracket ways. This research paper focused on the academic work regarding 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.[26] 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.

Jyoti and Prashant et al.[29] have categorized the events into anomalous or non-anomalous events using CNN. They use videos as input process them frame-wise and do this binary classification. However, they could not detect outliers in it.

Jyoti et al.[30] used the design of the deep learning model which was proposed for 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.

Swapnil and Sagar et al.[31] have proposed a deep learning approach to identify crime with the assistance of automated driving cars. They used object detection using OpenCV and classification using LSTM that results in tracking live streams like mob lynching or burglary. They have achieved 90

III. Methodology

A. Introduction & Dataset

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 Violence or Non-Violence. In preprocessing, maximum of 12 representative frames per video are extracted and resized to 299×299 pixels (the requirement for InceptionV3), and picture component values are normalized to the binary values, classified as 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 inputs 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.

B. Feature Extraction & Model Architectures

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 recognition of violent activities in the provided 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.

C. Training Methodology & Performance Evaluation

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

IV. Result

A. 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 real time video surveillance. 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 fighting from video sequences well.

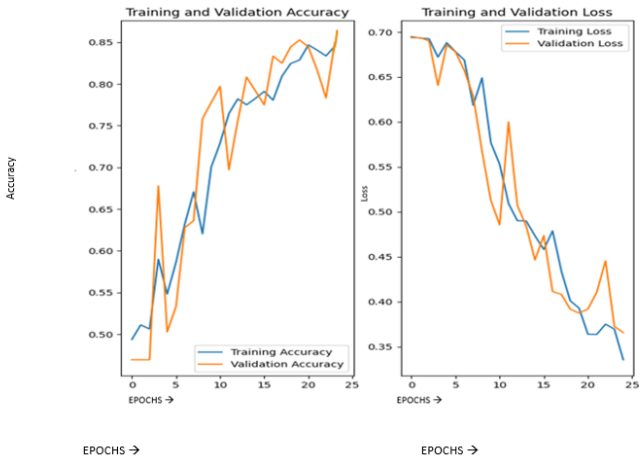

Fig.1 Performance of CNN for Fight Detection System (a) Accuracy (b) Loss variations with epochs

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.

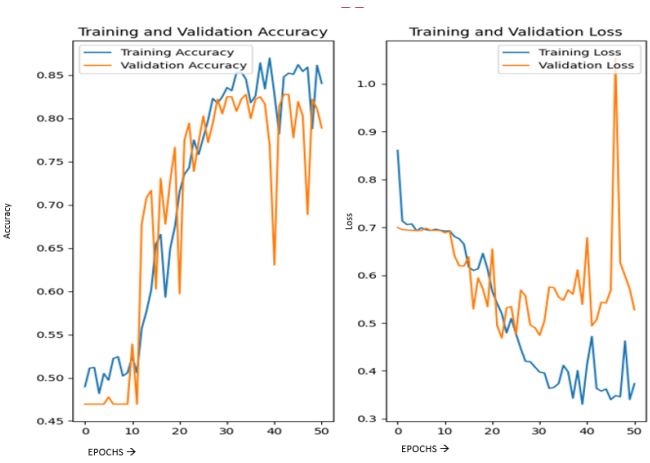

Fig.2 Performance of LSTM for Fight Detection System (a) Accuracy (b) Loss variations with epochs

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.

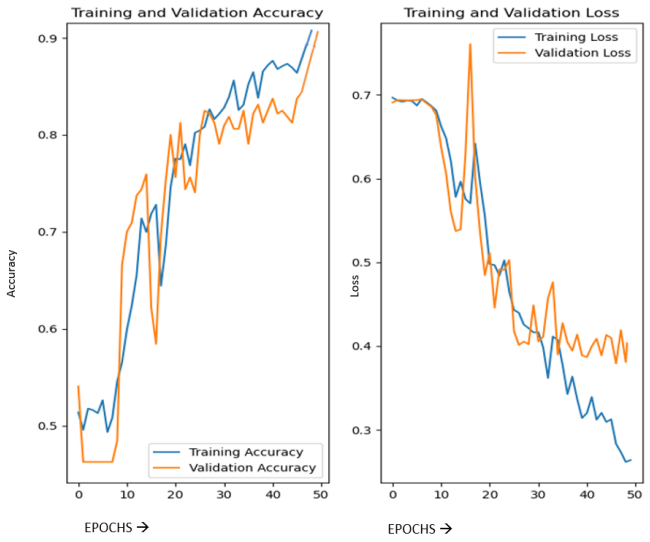

Fig.3 Performance of GRU for Fight Detection System (a) Accuracy (b) Loss variations with epochs

B. 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.

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.

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.

[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 AIRT-Lab 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.

[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] 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.

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

[11] 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.

[12] 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

[13] 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.

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

[15] 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.

[16] 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

[17] S.I.Sumon, "Human Activity Recognition using CNN and LSTM," GitHub Repository,2020.

[18] 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.

[19] 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.

[20] R. Halder and R. Chatterjee, "CNN-BiLSTM Model for Violence Detection in Smart Surveillance," SN Computer Science, vol. 1, 2020.

[21] 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).

[22] 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).

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

[24] 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.
[25] Kaur, Harjinder et al. “A Review of Machine Learning based Anomaly Detection Techniques.” ArXiv abs/1307.7286 (2013)
[26] Yao, Huiling and Xing Hu. “A survey of video violence detection.” Cyber-Physical Systems 9 (2021): 1 - 24.
[27] 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.
[28] European Union Agency for Fundamental Rights, Violence Against Women: An EU-Wide Survey – Main Results Report, 2014.
[29] 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.
[30] 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.
[31] 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.