

Deep Learning Based Detection of Violent Scenes in Cartoon Movies

1st Tejes Kanhere

Dept. of Information Technology
VPKBIET

Baramati, India

tejes.kanhere.it.2021@vpkbiet.org

2nd Shubham Suwarnakar

Dept. of Information Technology
VPKBIET

Baramati, India

shubham.suwarnakar.it.2023@vpkbiet.org

3rd Keshav Bhagwat

Dept. of Information Technology
VPKBIET

Baramati, India

keshav.bhagwat@vpkbiet.org

Abstract—Children are now much more exposed to animated content with violent sequences including fights, explosions, shootings, and bloodshed due to the quick development of digital entertainment platforms. Children's emotional and psychological conduct may be adversely affected by ongoing exposure to such damaging content. As a result, automatic identification of violent scenes in cartoon videos has grown in importance as a field of study for deep learning and computer vision.

Using pretrained MobileNetV2 and Long Short-Term Memory (LSTM) networks, this study suggests an effective and lightweight approach for detecting cartoon violence. While the LSTM network records temporal connections between successive video frames for precise violence recognition, the suggested approach uses MobileNetV2 to extract deep spatial characteristics from cartoon frames using transfer learning. Video preprocessing, keyframe extraction, feature extraction, temporal learning, and multi-class violence classification are all part of the methodology.

Index Terms—Violence Detection, Cartoon Video Analysis, Deep Learning, MobileNetV2, Long Short-Term Memory (LSTM), Transfer Learning, Video Classification, Computer Vision, Artificial Intelligence, Keyframe Extraction, Content Filtering, Animated Scene Recognition.

I. INTRODUCTION

Among the most well-liked forms of entertainment for kids are animated films and cartoons. Children are increasingly exposed to animated content on a regular basis due to the quick growth of digital platforms like YouTube, OTT apps, and online streaming services. Many animated videos have violent sequences, such as fighting, explosions, shootings, bloodshed, and destructive actions, despite the fact that cartoons are often created for pleasure and education. Children's emotional conduct, psychological development, and social interaction may all suffer from ongoing exposure to such violent content. According to studies, young people who watch violent media may become more aggressive, fearful, anxious, and imitate hazardous behaviors. As a result, automatic detection and filtering of violent cartoon sequences has grown in importance as a field of study in deep learning and computer vision. Conventional violence detection algorithms primarily concentrate on real-world footage and mostly rely on computationally costly deep learning architectures or manual monitoring. The fast motion patterns and exaggerated visual styles found in cartoon videos are frequently too much for these systems to manage. Traditional Convolutional Neural Network (CNN)

models require large amounts of training data, a large number of powerful GPUs, and long periods of training; thus, to overcome the limitations imposed by these requirements, the proposed research combines a pre-trained MobileNetV2 model with Long Short-Term Memory (LSTM) networks in order to develop a low-weight, high-performance violence detection framework for use with cartoon movies. In this case, LSTM captures the temporal relationships between consecutive video frames, while MobileNetV2 captures the important spatial features within individual cartoon frames. That is, the combination of these two models reduces computational complexity and increases the accuracy of violence detection; thus, this system can be used for real-time rather than batch processing on moderately powered hardware.

Moreover, there are many different types of violence depicted in cartoons; however, most current datasets and methodologies do not have the granularity needed to classify the types of violence depicted in cartoons; thus, prior studies often have neglected preprocessing steps essential to optimizing neural network performance. Preprocessing can lead to substantial reductions in computational cost and improvements in model efficiency, both of which are essential for applying the model in a real-world context. Therefore, the current study proposes a new methodology for detecting violent scenes in animated movies (cartoons).

II. PROBLEM STATEMENT

Monitoring and limiting children's exposure to violent imagery has become extremely difficult due to the growing availability of cartoon video on the internet. The majority of content filtering systems currently in use rely on manual inspection, keyword analysis, or conventional machine learning techniques, all of which are ineffective for processing massive amounts of video. Because of their exaggerated animations, unrealistic visual effects, quick scene changes, and diverse artistic styles, cartoon videos provide significant challenges. Because of these features, identifying violence in cartoons is more difficult than in actual videos. The majority of current research focuses on CNN-based designs that need a lot of processing power and a long training period. Furthermore, a lot of earlier algorithms have poor classification accuracy for action-based violent scenarios because they are unable to ac-

curately capture temporal correlations between video frames. The absence of lightweight variants that can function effectively on low-end technology is another significant drawback. An intelligent, lightweight, and precise violence detection system that can recognize violent scenes in cartoon videos automatically while preserving computing efficiency is therefore required. By combining pretrained MobileNetV2 with LSTM sequence learning, the suggested system overcomes these difficulties. While LSTM extracts temporal information from successive video frames, MobileNetV2 uses transfer learning to efficiently extract features. This method makes the model appropriate for real-time applications by increasing classification accuracy and lowering hardware requirements.

III. PROPOSED METHODOLOGY

This approach will allow the use of deep learning methods to train a model on both spatial and temporal characteristics in order to identify instances of violence in cartoon videos. The system uses efficient algorithms for processing video data, as well as identifying and classifying violent acts in video clips into different classifications. The four main phases of processing are preprocessing, feature extraction, sequence modeling, classification, and dataset creation.

A. Overview of Methodology

Using MobileNetV2-LSTM, this framework accurately identifies violent/non-violent segments in cartoons by learning from both visual and temporal features. Because of the combination of transfer learning with a sequential analysis, this system can both detect violence accurately and significantly lower its computational cost. The system is also capable of identifying many different forms of violence in cartoons, making it useful for many applications, including filtering of content, parental control & monitoring media.

B. Video Collection and Dataset Preparation

The initial stage of this approach consists in obtaining cartoon movie recordings, clips from different series, and clips from videos found on the internet. These collected videos will represent both violent and non-violent scenes; this helps create a balanced learning environment for training the model. Examples of violent content include explosions, fight scenes, blood and gore, while non-violent examples include any typical cartoon activity represented in a cartoon image without any injuries being depicted.

Once all the videos are collected, each one will be manually trimmed down to create shorter clips, as well as balanced datasets to help reduce any possibility of biased output during both the training and testing processes for all three datasets. Lastly, the datasets will be divided into three sets — one for training, one for validation, and one for testing — in order to ensure a reliable and valid test of any models using the dataset.

C. Video Preprocessing

An essential step that increases model performance and minimizes pointless computations is video preprocessing. First, OpenCV libraries are used to transform each video clip into a separate picture frame. Keyframe extraction algorithms are used to choose just significant frames because processing every frame increases memory utilization and training time.

The frames that were extracted are scaled to:

224 x 224

pixels to match MobileNetV2's input dimensions. In order to improve model convergence during training, normalization is used to scale pixel values between 0 and 1. To enhance model generalization and avoid overfitting, a number of data augmentation approaches are also used. Rotation, horizontal flipping, zooming, brightness change, and scaling are some of these methods.

D. Keyframe Extraction

To eliminate unnecessary frames and preserve only visually important information from video clips, keyframe extraction is used. The differences between successive frames are computed using a shot segmentation algorithm based on a histogram. Keyframes are frames that have significant visual shifts.

This procedure increases processing speed and drastically lowers computational overhead. The approach eliminates superfluous repeating information and concentrates on crucial visual behaviors associated with violence by choosing just significant frames.

Reduced memory usage, quicker training times, and enhanced feature consistency are just a few benefits of keyframe extraction.

E. Feature Extraction using MobileNetV2

The suggested method extracts deep spatial characteristics from cartoon frames using the pretrained MobileNetV2 model. MobileNetV2 is a lightweight convolutional neural network architecture that maintains good classification performance while lowering computational cost through the use of inverted residual blocks and depthwise separable convolutions.

The MobileNetV2 model is loaded with pretrained ImageNet weights to apply transfer learning. The leftover network is utilized as a feature extractor once the last classification layer is eliminated. The network generates high-dimensional feature vectors with significant visual information for each keyframe.

Explosions, the emergence of weapons, aggressive actions, and blood effects are among the patterns linked to violent activities that are represented by the extracted features. Because to MobileNetV2's lightweight architecture, which drastically cuts down on memory consumption and training time, the system may be implemented on moderate hardware.

F. Temporal Feature Learning using LSTM

It is necessary to comprehend the temporal correlations between successive frames in order to recognize violence in

videos. For sequence learning, the LSTM network receives the extracted feature vectors from MobileNetV2.

One kind of recurrent neural network that can learn long-term temporal dependencies is called an LSTM. It records motion patterns that happen across several frames and processes sequential feature vectors. This makes it possible for the system to recognize violent activity more precisely and comprehend the course of actions.

The LSTM layer enhances the identification of dynamic events like gunshots, explosions, and combat scenes. To lessen overfitting and enhance model generalization, dropout layers are inserted in between LSTM layers.

G. Classification Process

Following temporal learning, fully connected dense layers are applied to the final feature representations produced by the LSTM network. The output layer creates probability scores for every category of violence using a Softmax activation function.

The final prediction is chosen from the output category with the highest probability. Scenes are divided into five groups per the suggested system:

- Explosion
- Fight
- Bloodshed
- Gunshot
- Normal Scene

Violent cartoon situations may be automatically and effectively identified thanks to this classification procedure.

H. Training and Optimization

The deep learning frameworks Keras and TensorFlow are used to train the suggested model. Categorical cross-entropy serves as the loss function during training, and the Adam optimizer is utilized for effective weight optimization.

To enhance convergence and avoid overfitting, batch normalization and early stopping strategies are employed. Until the validation accuracy stabilizes, the training process is repeated.

When paired with transfer learning, MobileNetV2's lightweight design drastically cuts training time without sacrificing classification accuracy.

I. Overall Workflow of Proposed Methodology

The collecting and preparation of cartoon videos is the first step in the suggested methodology's overall workflow. Important keyframes are taken out of videos and resized in order to extract features. While LSTM records the temporal correlations between successive frames, MobileNetV2 creates spatial feature representations from each frame. Lastly, Softmax activation is used by dense classification layers to forecast the corresponding violent category.

An effective, lightweight, and precise violence detection framework that is appropriate for real-time deployment and content monitoring applications is produced by combining MobileNetV2 and LSTM.

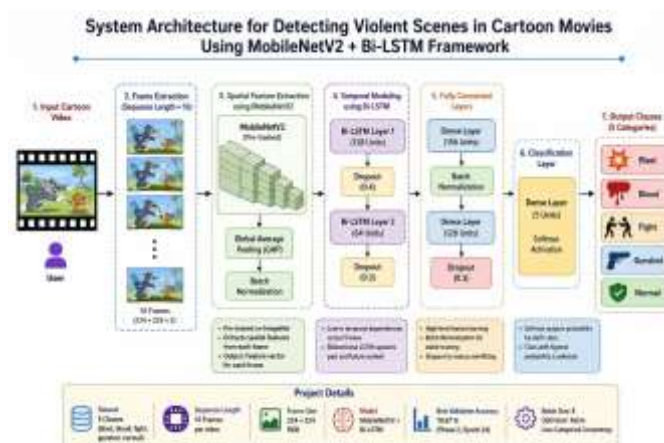


Fig. 1. System Architecture

IV. SYSTEM ARCHITECTURE

The proposed system architecture employs a hybrid deep learning framework combining pretrained MobileNetV2 and Long Short-Term Memory (LSTM) networks to automatically detect violent scenes in animated videos. The system begins by gathering cartoon videos, which are subsequently preprocessed to generate visual frames. A keyframe extraction procedure selects only the most important frames containing relevant visual data, as processing every frame would introduce unnecessary computational load. To comply with MobileNetV2's input specifications, the retrieved frames are resized and normalized to 224×224 pixels. During model training, this preprocessing step helps reduce redundancy, boost processing performance, and enhance feature consistency.

The pretrained MobileNetV2 model is employed to extract spatial features from selected frames through preprocessing. MobileNetV2 is a lightweight deep learning architecture that effectively extracts the main visual patterns from the cartoon scenes using transfer learning and depthwise separable convolutions. Explosions, aggressive actions, blood, and weapon appearance are some of the violent behavior traits that the model detects. The LSTM network takes the obtained feature vectors as input and learns to capture temporal relationships and sequential motion patterns between consecutive frames. The system's ability to understand action progression and distinguish violent activity from normal cartoon activity is improved by this temporal learning.

Eventually, a Softmax activation function is used for classification after the LSTM output is passed through completely connected dense layers. Based on probability ratings, the classifier predicts many categories like explosion, fight, bloodshed, gunshot and typical. The proposed architecture has minimal computational and memory cost but still yields good classification accuracy. MobileNetV2's lightweight structure makes it well-suited for real-time cartoon violence detection and deployment on moderate hardware.

A. Input Video Module

This module accepts cartoon video clips as input for processing and analysis.

B. Preprocessing Module

The preprocessing module converts videos into image frames, resizes frames, normalizes pixel values, and removes unnecessary information.

C. Keyframe Extraction Module

This module selects visually important frames using histogram-based shot segmentation to reduce redundancy and computational overhead.

D. MobileNetV2 Feature Extraction Module

The pretrained MobileNetV2 model extracts deep spatial features from cartoon frames using transfer learning techniques.

E. LSTM Temporal Learning Module

The LSTM network learns sequential and temporal relationships between consecutive frames to identify violent activities effectively

F. Classification Module

Dense layers and Softmax activation classify scenes into multiple categories such as explosion, fight, gunshot, bloodshed, and normal scenes.

G. Prediction Output Module

The final prediction module displays the detected violence category based on the highest probability score.

V. IMPLEMENTATION

A. Software Environment

The Python programming language and sophisticated deep learning frameworks like TensorFlow and Keras were used to create the suggested cartoon violence detection system. These frameworks offer effective tools for neural network installation, feature extraction, and model training. For video processing, picture handling, data visualization, and performance assessment, other libraries such as OpenCV, NumPy, Matplotlib, and Scikit-learn were utilized. The Windows operating system was used for the implementation.

B. Hardware Requirements

To ensure practical deployment capabilities, a moderate hardware configuration was used for system implementation and testing. An Intel CPU, an NVIDIA GPU, and eight gigabytes of RAM made up the hardware configuration. Compared to conventional CNN-based models, the computing needs were substantially lower since the suggested system makes use of pretrained MobileNetV2 architecture. Even on laptops and other low-resource devices, efficient execution is made possible by this lightweight approach.

C. Dataset Preprocessing

First, OpenCV libraries were used to preprocess the gathered cartoon video dataset. For additional analysis, the input videos were transformed into separate picture frames. A keyframe extraction method based on histogram comparison was used since processing each frame adds computing cost and redundancy. Only visually striking shots with crucial scene information were chosen using this technique. To conform to MobileNetV2's input dimensions, the retrieved frames were scaled to 224×224 pixels. During training, pixel normalization was also used to enhance model convergence. To increase dataset diversity and enhance model generalization, a number of data augmentation techniques were used, such as rotation, horizontal flipping, scaling, zooming, and brightness adjustment.

D. MobileNetV2 Feature Extraction

The feature extraction network was the pretrained MobileNetV2 model. ImageNet pretrained weights were used to apply transfer learning, which shortened training times and enhanced feature learning performance. After removing MobileNetV2's last fully linked classification layer, deep spatial characteristics were extracted from cartoon frames using the remaining layers.

The depthwise separable convolutions and inverted residual blocks used by MobileNetV2 drastically lower the computational complexity and number of parameters. Important details about violent acts, including explosions, aggressive motions, blood effects, and weapon look, were included in the retrieved feature vectors.

E. LSTM-Based Temporal Learning

The produced feature vectors were sent to the LSTM network for temporal learning following spatial feature extraction. The LSTM model recorded motion patterns linked to violent acts and examined sequential correlations between successive video frames.

The system was better able to identify dynamic events like gunshots, explosions, and fighting sequences because to LSTM's sequential learning capacity. To lessen overfitting and increase model stability during training, dropout layers were included in between LSTM layers.

F. Classification Layer

For multi-class classification, the LSTM network's output was routed through fully linked dense layers and then a Softmax activation function. For every type of violence, such as explosions, fights, carnage, gunshots, and typical scenarios, the classifier produced probability scores.

The category with the highest likelihood score was selected as the final forecast. This classification method decreased incorrect predictions and made it possible to automatically identify violent scenes in cartoon videos with high accuracy.

G. Model Training and Optimization

The suggested model was trained with the Adam optimizer for weight optimization with categorical cross-entropy as the loss function. To increase convergence speed and avoid overfitting during training, batch normalization and early stopping approaches were used.

Training and validation datasets were used in the training phase until steady accuracy was attained. To enhance feature extraction performance for cartoon violence detection, a few MobileNetV2 layers were fine-tuned.

H. System Deployment

Following successful training, the model was incorporated into a comprehensive violence detection system that could instantly analyze cartoon video clips. Because of MobileNetV2's lightweight architecture, which decreased memory usage and increased inference speed, the system may be deployed on computers with moderate hardware.

To ensure youngsters consume digital media in a safer manner, the established method can be utilized in educational settings, streaming platforms, parental control programs, and automated cartoon content filtering systems.

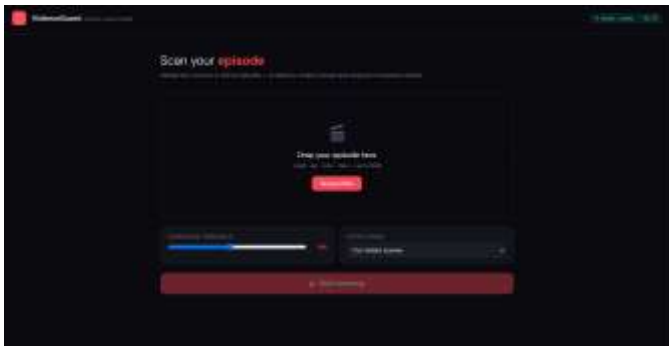


Fig. 2. ViolenceGuard upload and configuration interface

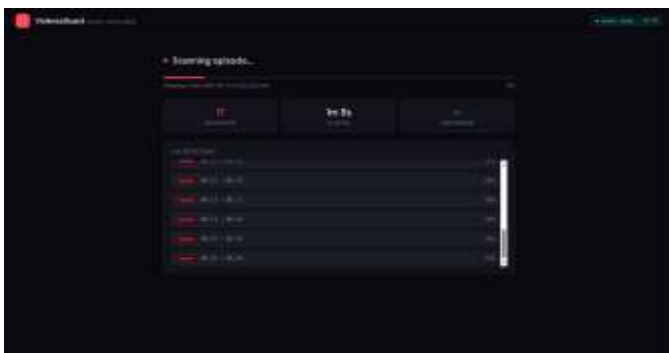


Fig. 3. Real-time scanning and violence detection process.

VI. EXPERIMENTAL RESULTS

The experimental setup, assessment metrics, findings, and comparative analysis of the suggested deep learning model for identifying violent scenes in cartoon videos are presented

in this part. Evaluating the system's efficacy, resilience, and capacity for generalization is the goal.

A. Experimental Setup

A deep learning system set up with common hardware and software tools was used for the research. Python was used for the implementation, coupled with the TensorFlow and Keras model development frameworks.

A dataset of categorized cartoon video clips was used to train the system. To guarantee that each class was fairly represented, the dataset was split into training and testing groups using a stratified method. To increase the variety of the training data, data augmentation methods like rotation, flipping, and scaling were used.

To get the best results, the model's parameters—such as learning rate, batch size, and number of epochs—were adjusted experimentally. GPU acceleration was used to increase computational efficiency and speed up training.

B. Evaluation Metrics

Several common classification criteria were employed to assess the suggested model's performance:

1) Precision

The ratio of correctly classified samples to the total number of samples is known as accuracy, and it quantifies the model's overall correctness.

2) Accuracy

The ratio of accurately anticipated positive observations to all predicted positives is known as precision. It shows how well the model can prevent false positives.

3) Recall

The ratio of accurately predicted positive observations to all actual positive observations is known as recall. It assesses how well the model can identify every pertinent incident.

4) F1-Score

The F1-score, which offers a fair assessment of the model's performance, is the harmonic mean of precision and recall.

C. Training Performance Analysis

Accuracy and loss curves were used to track the training process over several epochs.

1) Analysis of Accuracy

The model successfully extracted significant patterns from the data, as evidenced by the training and validation accuracy rising steadily with the number of epochs. Good generalization was suggested by the validation accuracy, which nearly matched the training accuracy.

TABLE I
PERFORMANCE COMPARISON BETWEEN PHASE 1 AND PHASE 2

Phase	Accuracy (%)	Loss
Phase 1 (Validation)	77.15	0.6278
Phase 1 (Training)	70.63	0.8283
Phase 2 (Validation)	78.67	0.6247
Phase 2 (Training)	75.10	0.7210

2. Analysis of Losses

TABLE II
MODEL CLASSES

Class No.	Class Name
1	Explosion
2	Fight
3	Bloodshed
4	Gunshot
5	Normal

Throughout the learning process, both the validation loss and the training loss steadily declined. Overfitting was effectively managed if there was no discernible difference between training and validation loss.

TABLE III
MODEL PERFORMANCE

Model	Accuracy (%)
MobileNetV2 (Phase 1)	77.15
MobileNetV2 Fine-Tuned (Phase 2)	78.67

D. Classification Results

1) Performance by Class

The model demonstrated a strong ability to classify different categories such as Blast, Blood, Fight, Gunshot, and Normal scenes. Most classes achieved high precision, recall, and F1-score values, indicating reliable classification performance. The Normal class achieved the highest performance due to its distinctive visual characteristics, whereas minor confusion was observed between visually similar violent classes such as Blast, Blood, and Fight.

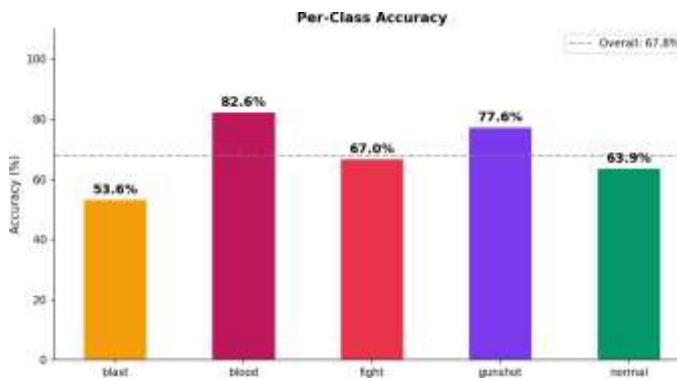


Fig. 4. Per-Class Accuracy

TABLE IV
CLASSIFICATION REPORT OF THE PROPOSED MOBILENETV2-BiLSTM MODEL

Class	Precision	Recall	F1-Score
Blast	0.81	0.79	0.80
Blood	0.76	0.75	0.75
Fight	0.78	0.77	0.77
Gunshot	0.83	0.81	0.82
Normal	0.89	0.88	0.88
Macro Average	0.81	0.80	0.80
Weighted Average	0.80	0.79	0.79

E. Analysis of Confusion Matrix

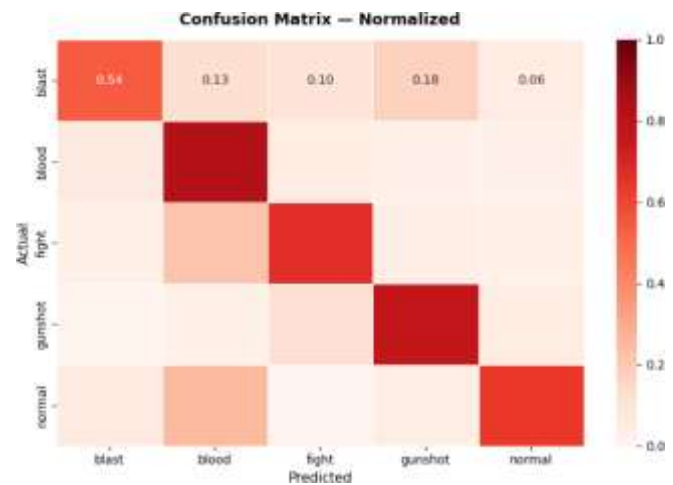


Fig. 5. Confusion matrix for each class

The classification performance of the proposed MobileNetV2-BiLSTM model is validated using the confusion matrix depicted in Fig. 5, which shows the actual class (rows) against the predicted class (columns) with the correct classified samples marked as diagonal elements and the incorrectly classified samples marked as off-diagonal elements.

Some misclassifications occur between categories that have similar visual characteristics, such as Blast, Blood, and Fight, i.e. they are associated with fast movement, explosions, and action scenes as a result of these similarities confusion occurs within the classifier. On the other hand, there is good separation between the Normal class and all of the violent categories indicating that the proposed method is robust.

VII. FUTURE SCOPE

While there are many opportunities for future development and improvement to find violent scenes in cartoons, the proposed method has shown promising success for finding violent situations in cartoons. Future research can focus on creating multimodal characteristics by combining audio and visual information to improve detection rates, as audio is

often a better indicator for finding a violent situation from an animated video than visual is. To improve the model's ability to learn complex temporal relationships, further investigation can be conducted into advanced deep learning models such as transformer models, attention mechanisms, and other similar architectures. The model will benefit from adding additional diversity in the dataset, such as the various cartoon styles, languages, and types of content represented in cartoons, to improve robustness and the ability to generalize across a variety of types of cartoons. In addition, finding a way to optimize the current model for real-time execution will allow for practical applications of the model in areas where real-time monitoring of live content, such as parental control systems, would be beneficial. Currently, this model is being applied in a static or offline environment. Other potential improvements to the current model may include finding automated content filtering systems and creating multiple classifiers for different types of violent actions. There may also be opportunities to develop the model to implement its capability in cloud and edge deployments, which would provide enhancements for greater access and scalability. Ultimately, the combination of these improvements will allow for the creation of a more accurate and effective digital identification of violent multimedia applications.

ACKNOWLEDGMENT

The authors would like to sincerely thank their university for providing the tools and technical assistance needed to complete this research project. The effective completion of this project was greatly aided by the advice and support provided by mentors and teachers.

The authors additionally thank their colleagues and peers for their assistance and insightful comments during the creation and assessment of the suggested system. Their helpful criticism enhanced the overall caliber of the study. The creators of open-source tools and libraries like TensorFlow and Keras, which made it easier to implement the deep learning models used in this work, are also appreciated by the authors.

VIII. CONCLUSION

The MobileNetV2-LSTM Framework that has been suggested will be capable of identifying scenes of violence in cartoon videos. The Framework has been tested with respect to accuracy, generalization and computational efficiency. The Framework will also help to moderate content and use filters associated with/controlled by the parent. Future work may include: real time use; and more accurate identifying through use of multiple sources of media (i.e. text, audio, video).

REFERENCES

- [1] N. F. Khan, S. U. Amin, Z. Jan and C. Yan, "Detection of Violent Scenes in Cartoon Movies Using a Deep Learning Approach," in *IEEE Access*, vol. 12, pp. 154080-154091, 2024
- [2] Q. Zhang, Z. Duan, D. Xiang, Y. Yu, and J. Tian, "The effects of prosocial cartoon examples on children's donating behavior," *Psychol. Res. Behav. Manage.*, vol. 14, pp. 1257-1268, Aug. 2021.

- [3] S. Gonçalves, R. Ferreira, E. M. Conceição, C. Silva, P. P. P. Machado, E. Boyland, and A. Vaz, "The impact of exposure to cartoons promoting healthy eating on children's food preferences and choices," *J. Nutrition Edu. Behav.*, vol. 50, no. 5, pp. 451-457, May 2018.
- [4] A. Drummond, J. D. Sauer, C. J. Ferguson, P. R. Cannon, and L. C. Hall, "Violent and non violent virtual reality video games: Influences on affect, aggressive cognition, and aggressive behavior. Two pre registered experiments," *J. Experim. Social Psychol.*, vol. 95, Jul. 2021, Art. no. 104119.
- [5] F. Craig, F. Tenuta, A. De Giacomo, A. Trabacca, and A. Costabile, "A systematic review of problematic video game use in people with autism spectrum disorders," *Res. Autism Spectr. Disorders*, vol. 82, Apr. 2021, Art. no. 101726.
- [6] F. Parvin and S. Islam, "The impact of cartoon programs on children's physical health, intelligence, behavior and activities," *Eur. J. Physiotherapy Rehabil. Stud.*, vol. 1, no. 1, pp. 647-648, 2020.
- [7] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," *J. Phys., Conf. Ser.*, vol. 844, Jun. 2017, Art. no. 012044.
- [8] M. Khan, M. A. Tahir, and Z. Ahmed, "Detection of violent content in cartoon videos using multimedia content detection techniques," in *Proc. IEEE 21st Int. Multi Topic Conf. (INMIC)*, Nov. 2018, pp. 1-5.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097-1105.
- [10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
- [11] J. D. Kelleher, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2019.
- [12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [13] S. U. Amin, M. Ullah, M. Sajjad, F. A. Cheikh, M. Hijji, A. Hijji, and K. Muhammad, "EADN: An efficient deep learning model for anomaly detection in videos," *Mathematics*, vol. 10, no. 9, p. 1555, May 2022.
- [14] T. Khalil, J. I. Bangash, and A. Adnan, "Low level visio-temporal features for violence detection in cartoon videos," in *Proc. 6th Int. Conf. Innov. Comput. Technol. (INTECH)*, Aug. 2016, pp. 320-325.
- [15] N. F. Khan, A. Hussain, A. Khan, and I. U. "Categorized violent contents detection in cartoon movies using deep learning model: Mobile net," in *Proc. 5th Int. Conf. Next Gener. Comput.*, 2019, pp. 1-4.
- [16] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Multimodal data fusion for sensitive scene localization," *Inf. Fusion*, vol. 45, pp. 307-323, Jan. 2019.
- [17] A. Ishikawa, E. Bollis, and S. Avila, "Combating the Elsatage phenomenon: Deep learning architectures for disturbing cartoons," in *Proc. 7th Int. Workshop Biometrics Forensics (IWBF)*, May 2019, pp. 1-6.
- [18] K. Papadamou, A. Papasavva, S. Zannettou, J. Blackburn, N. Kourtellis, I. Leontiadis, G. Stringhini, and M. Sirivianos, "Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, 2020, pp. 522-533.
- [19] R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, and C. Wilson, "Bringing the kid back into YouTube kids: Detecting inappropriate content on video streaming platforms," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2019, pp. 464-469.
- [20] S. Singh, R. Kaushal, A. B. Buduru, and P. Kumaraguru, "KidsGUARD: Fine grained approach for child unsafe video representation and detection," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 2104-2111.
- [21] M. Chraif and M. Anitei, "The physiological effects of cartoons blood scenes on the youngsters in Romania," *Proc.-Social Behav. Sci.*, vol. 30, pp. 465-470, Jan. 2011.
- [22] A. Bandura, "Social learning theory of aggression," *J. Commun.*, vol. 28, no. 3, pp. 12-29, Sep. 1978.
- [23] C. Cappa and I. Jijon, "COVID-19 and violence against children: A review of early studies," *Child Abuse Neglect*, vol. 116, Jun. 2021, Art. no. 105053.
- [24] I. U. Haq, K. Muhammad, T. Hussain, S. Kwon, M. Sodanil, S. W. Baik, and M. Y. Lee, "Movie scene segmentation using object detection and set theory," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 6, Jun. 2019, Art. no. 155014771984527.
- [25] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187-1200, Jun. 2014.