International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

AUTOMATED INAPPROPRIATE CONTENT DETECTION ON YOUTUBE: A DEEP LEARNING FRAMEWORK

[1]M. Mounika, [2]B.Aparna
[1]Assistant Professor, Depatment of MCA, Sree Chaitanya College of Engineering, Karimnagar
[2]MCA Student, Depatment of MCA, Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT

YouTube's video content has grown exponentially, drawing billions of viewers, the vast majority of whom are young. Additionally, malicious uploaders use this site to disseminate disturbing visual content, such as improper content for minors via animated cartoon videos. Therefore, it is strongly advised that social media networks have an automated real-time video content screening method. This paper suggests a revolutionary deep learning-based architecture for identifying and categorising films that include unsuitable information. A pre-trained convolutional neural network (CNN) model called EfficientNet-B7 from ImageNet is used in the suggested framework to extract video descriptors. These are subsequently fed into a bidirectional long short-term memory (BiLSTM) network, which learns efficient video representations and performs multiclass video classification. After BiLSTM, an attention mechanism is also included to apply the network's attention probability distribution. A carefully annotated dataset of 111,156 cartoon clips gathered from YouTube videos is used to assess these algorithms. According to experimental data, the EfficientNet-BiLSTM framework (accuracy D 95.66%) outperforms the attention mechanism-based Efficient Net-Bi LSTM framework (accuracy D 95.30%). Second, deep learning classifiers outperform typical machine learning classifiers in terms of performance. All things considered, the Efficient Net and Bi LSTM design with 128 hidden units produced cutting-edge results (f1 score D 0.9267). Furthermore, the performance comparison against current state-of-the-art methods confirmed that BiLSTM on top of CNN achieves better results in the detection and classification of child inappropriate video content because it captures better contextual information of video descriptors in network architecture.

A key element of the expanding discipline of data science is machine learning. Various algorithms are taught using statistical techniques to create predictions or classifications and to provide important information for this project. These insights then inform business and application decision-making, hopefully influencing important growth indicators.

This project data, sometimes referred to as training data, is used by machine learning algorithms to create a model that allows them to make judgements or predictions without explicit programming. When developing traditional algorithms to accomplish the required tasks is challenging or impractical, machine learning methods are employed in a wide range of datasets.

1. INTRODUCTION

The creation and consumption of videos on social media platforms have grown drastically over the past few years. Among the social media sites, YouTube predominates as a video sharing platform with plethora of videos from diverse categories. According to YouTube statistics [1], the global user base of YouTube is over 2 billion registered users and more than 500 hours of video content is uploaded every minute. Consequently, billions of hours of videos are available where users of all age groups can explore generic as well as personalized content [2]. Considering such a large-scale The associate editor coordinating the review of this manuscript and approving it for publication was Aasia Khanum . crowdsourced database, it is extremely challenging to monitor and regulate the uploaded content as per platform guidelines. This creates opportunities for malicious users to indulge in spamming activities by misleading the audiences with falsely advertised content (i.e., video, audio or text). The most disruptive behavior by malicious users is to expose the young audiences to disturbing content, particularly when it is fabricated as safe for them. Children today spend most of their time on the Internet and the YouTube platform for them has distinctly established itself as an alternative to traditional screen media (e.g., television) [3], [4]. The YouTube press release [5] also confirmed the high popularity of this social media site among younger audiences compared to other age groups, and the reason for this high level of approval is due to fewer restrictions [6].

Unlike television, children can be presented with any type of content on the Internet due to lack of regulations. Exposing children to disturbing content is considered as one among other internet safety threats (like cyberbullying, cyber predators, hate etc.) [7]. Bushman and Huesmann [8] confirmed that frequent exposure to disturbing video content may have a short-term or long-term impact on children's behavior, emotions and cognition. Many reports [9]–[12] identified the trend of distributing inappropriate content in children's videos. This trend got people's attention when mainstream media reported about the Elsagate controversy [13], [14], where such video material was found on YouTube featuring famous childhood cartoon characters (i.e., Disney characters, superheroes, etc.) portrayed in disturbing scenes; for instance, performing mild violence, stealing, drinking alcohol and involving in nudity or sexual activities. In an attempt to provide a safe online platform, laws like the children's online privacy protection act (COPPA) imposes certain requirements on websites to adopt safety mechanisms for children under the age of 13. YouTube has also included a ''safety mode'' option to filter out unsafe content. Apart from that, YouTube developed the YouTube Kids application to allow parental control over videos that are approved as safe for a certain age group of children [15]. Regardless of YouTube's efforts in controlling the unsafe content phenomena, disturbing videos still appear [16]–[19] even in YouTube Kids [20] due to difficulty in identifying such content. An explanation for this may be that the rate at which videos are uploaded every minute makes YouTube vulnerable to unwanted content. Besides, the decision-making algorithms of YouTube rely heavily on the metadata of video (i.e., video title, video description, view count, rating, tags, comments, and community flags).

Hence, filtering videos based on the metadata and community flagging is not sufficient to assure the safety of children [21]. Many cases exist on YouTube where safe video titles and thumbnails are used for disturbing content to trick children and their parents. The sparse inclusion of child inappropriate content in videos is another common technique followed by malicious uploaders. Fig. 1 displays an example among such cases where video title and video clips are safe for children (as shown in Fig. 1(a)) but included inappropriate scenes in this video (as shown in Fig. 1(b) and Fig. 1(c)). The concerning thing about this example, including many similar cases, is that these videos have millions of views with more likes than dislikes, and have been available for years. Many other cases (as shown in Fig. 1(d)) also identified where videos or the YouTube channel is not popular, yet contains child unsafe content especially in the form of animated cartoons. It is evident from examples that this problem persists irrespective of channel or video popularity. Furthermore, YouTube has disabled the dislike feature of videos which resulted in viewers being incapable of getting the indirect video content feedback from statistics. Since the YouTube metadata can be easily manipulated, it is suggested to better use video features for detection of inappropriate content than metadata features associated with videos [22].

2. EXISTINGSYSTEM

Rea *et al.* [37] proposed a periodicity-based audio feature extraction method which was later combined with visual features for illicit content detection in videos.

The machine learning algorithms are usually employed as classifiers Liu *et al.* [38] classified the periodicity-based audio and visual segmentation features through support vector machine (SVM) algorithm with Gaussian radial basis function (RBF) kernel. Later on, they extended the framework [39] by applying the energy envelope (EE) and bag-of-words (BoW)-based audio representations and visual features.

Ulges *et al.* [23] used MPEG motion vectors and Mel-frequency cepstral coefficient (MFCC) audio features with skin color and visual words. Each feature representation is processed through an individual SVM classifier and combined in a weighted sum of late fusion. Ochoa *et al.* [40] performed binary video genre classification for adult content detection by processing the spatiotemporal features with two types of SVM algorithms: sequential minimal optimization (SMO) and LibSVM.

Jung *et al.* [41] worked with the one dimensional signal of spatiotemporal motion trajectory and skin color. Tang *et al.* [42] proposed a pornography detection system_PornProbe, based on a hierarchical latent Dirichlet allocation (LDA) and SVM algorithm. This system combined an unsupervised clustering in LDA and supervised learning in SVM, and achieved high efficiency than a single SVM classifier. Lee *et al.* [43] presented a multilevel hierarchical framework by taking the multiple features of different temporal domains. Lopes *et al.* [44] worked with the bag-of-visual features (BoVF) for obscenity detection.

Kaushal *et al.* [21] performed supervised learning to identify the child unsafe content and content uploaders by feeding the machine learning classifiers (i.e., random forest, K-nearest neighbor, and decision tree) with video-level, user-level and comment-level metadata of YouTube Reddy *et al.* [45] handled the explicit content problem of videos through text classification of YouTube comments. They applied bigram collocation and fed the features to the naïve Bayes classifier for final classification.

Disadvantages

- An existing system doesn't ANALYSIS OF PRE-TRAINED CNN MODEL VARIANTS.
- An existing system doesn't ANALYSIS OF EFFICIENT-NET FEATURES WITH DIFFERENT CLASSIFIER VARIANTS.

3. PROPOSED SYSTEM

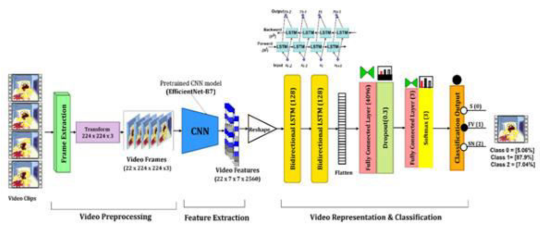
1. The system proposes a novel CNN (EfficientNet-B7) and BiLSTM-based deep learning framework for inappropriate video content detection and classification.
2. The system presents a manually annotated ground truth video dataset of 1860 minutes (111,561 seconds) of cartoon videos for young children (under the age of 13). All videos are collected from YouTube using famous cartoon names as search keywords. Each video clip is annotated for either safe or unsafe class. For the unsafe category, fantasy violence and sexual-nudity explicit content are monitored in videos. We also intend to make this dataset publicly available for the research community.
3. The system evaluates the performance of our proposed CNN-BiLSTM framework. Our multiclass video classifier achieved the validation accuracy of 95.66%. Several other state-of-the-art machine learning and deep learning architectures are also evaluated and compared for the task of inappropriate video content detection.

Advantages

- The most frequent applications of image/video classification employed the convolutional neural networks.
- The EfficientNet model is a convolutional neural network model and scaling method that uniformly scales network depth, width and resolution through compound co efficient.

4. SYSTEM ARCHITECTURE

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org



5. MODULES

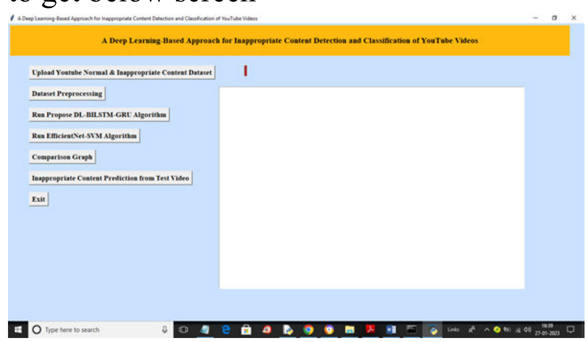
To implement this project we have designed following modules

1) **Upload YouTube Normal & Inappropriate Content Dataset:** using this module we will upload YouTube dataset images to application
2) Dataset Preprocessing: using this module we will read all images and then resize all images to equal size and then normalize image pixel values and then shuffle the dataset
3) **Run Propose DL-BILSTM-GRU Algorithm:** using this module we will split dataset into train and test and then input 80% training data to Pre-Trained CNN (EfficientNetB7) algorithm to extract digital content from images and then those features will get retrained with BI-LSTM algorithm to train a model. Trained model will be applied on 20% test data to calculate prediction accuracy
4) **Run EfficientNet-SVM Algorithm:** EfficientNetB7 features will get retrained with existing SVM algorithm and then calculate prediction accuracy
5) **Comparison Graph:** using this module we will plot accuracy comparison graph between propose EfficientNetB7-BILSTM and EfficientNetB7-SVM.
6) **Inappropriate Content Prediction from Test Video:** using this module we will upload any YouTube and if video contains fighting or violence then application will predict as 'Inappropriate Content' otherwise will predict SAFE content.

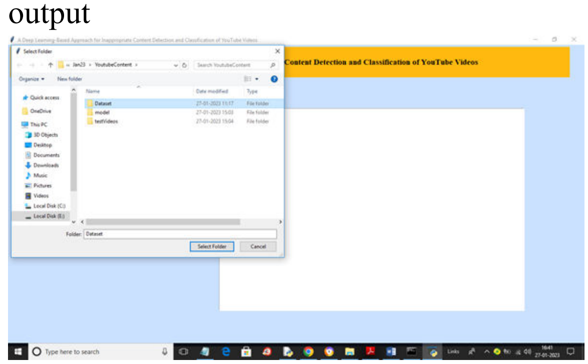
6. SCREEN SHOTS

In above screen we have two folders and juts go inside any folder to view training images

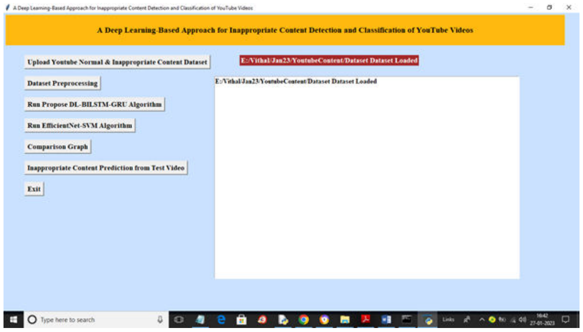
To run project double click on 'run.bat' file to get below screen



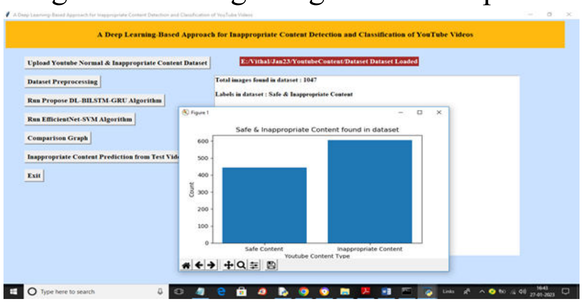
In above screen click on 'Upload YouTube Normal & Inappropriate Content Dataset' button to upload dataset and get below output



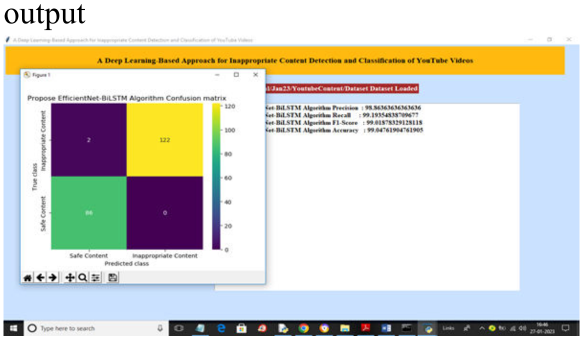
In above screen selecting and uploading entire ‘Dataset’ folder and then click on ‘Select Folder’ button to load dataset and get below output



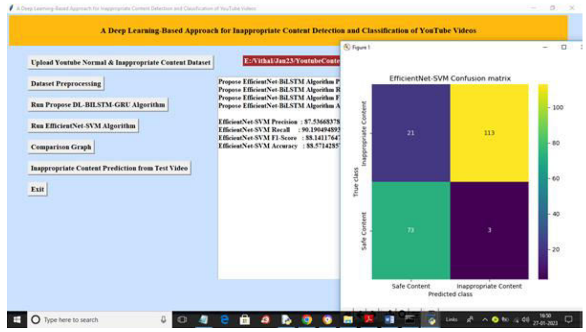
In above screen dataset loaded and now click on ‘Dataset Preprocessing’ button to read all images and then processes those images for training and get below output



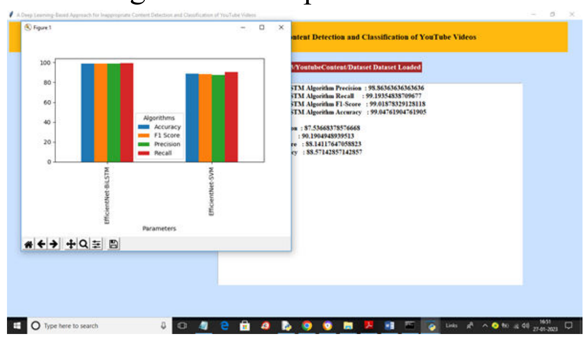
In above screen we can see dataset contains 1047 images and then in graph x-axis represents type of images such as ‘Safe and Inappropriate’ and y-axis represents count of those images. Now dataset processing completed and now click on ‘Run Propose DL-BILSTM-GRU Algorithm’ button to train propose algorithm and get below output



In above screen with propose EfficientNetB7-BI-LSTM we got 99.04% accuracy and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and green and yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count which is 2 only. Now close above graph and then click on ‘Run EfficientNet-SVM Algorithm’ button to get below output



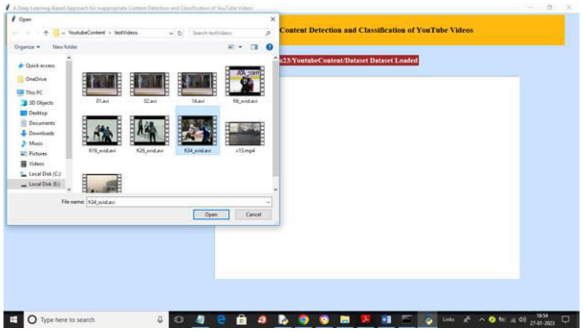
In above screen with EfficientNetB7-SVM we got 88% accuracy and in confusion matrix graph we can see in blue boxes that SVM predicted total 24 incorrect prediction so its accuracy is less. Now close above graph and then click on ‘Comparison Graph’ button to get below output



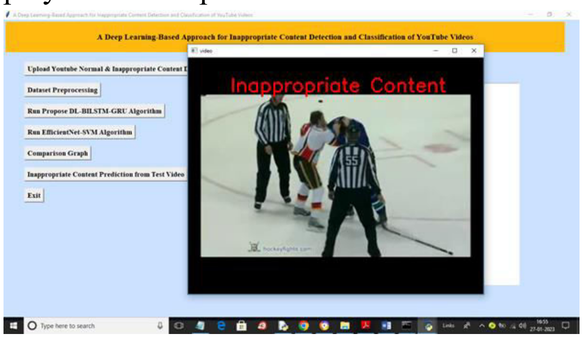
In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. In both algorithms propose EfficientNetB7-BI-LSTM got high accuracy. Now close above graph and then click on ‘Inappropriate Content Prediction from Test Video’ button

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

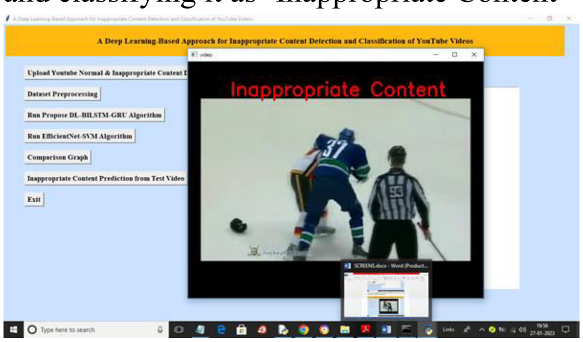
to upload test video and classify it as Safe or inappropriate.



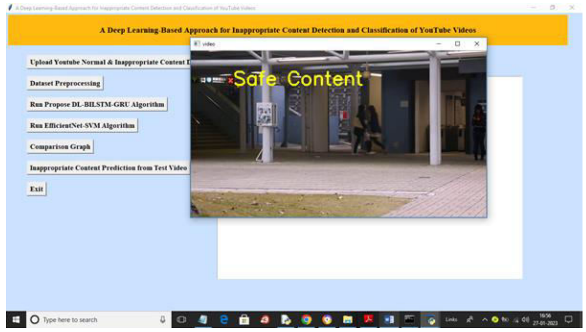
In above screen selecting and uploading video and then click on "Open' button to play video and perform classification



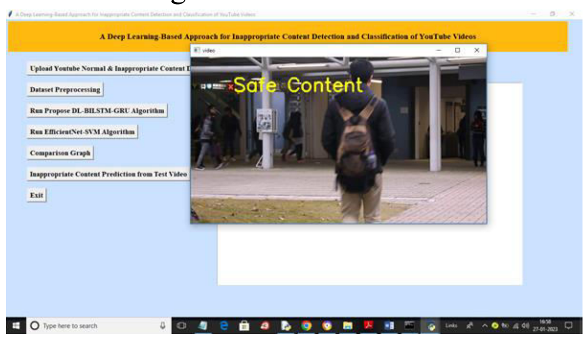
In above screen propose algorithm evaluating playing video and then detecting and classifying it as 'Inappropriate Content'



In above video also we got classification output



In above we got result as Safe Content



In above screen we got output as Safe Content as peoples are only moving in the video.

7. CONCLUSION

For the identification and categorisation of child-inappropriate video material, a unique deep learning-based system is put forward. The efficientnet-b7 architecture is used for transfer learning in order to extract video features.

The bilstm network processes the retrieved video features, learning the most effective video representations and performing multiclass video classification. A carefully annotated dataset of 111,156 cartoon video clips gathered from YouTube is used for all assessment studies. Compared to other tested models, such as efficient net-fc, efficient net-svm, efficient net-knn, efficient net-random forest, and efficient net-bilstm with attention mechanism-based models (with hidden units = 64, 128, 256, and 512), the evaluation results showed that the suggested framework of efficient net bilstm (with hidden units = 128) performs better (accuracy = 95.66%). Furthermore, by attaining the maximum recall score of 92.22%, the performance comparison with current state-of-the-art models also showed that our bilstm-based framework outperformed other models and approaches. The following are the benefits of the suggested deep learning-based approach for identifying unsuitable video content for kids:

1) The efficientnet-b7 and bilstm-based deep learning framework is used to process the video at a rate of 22 frames per second while taking into account the real-time circumstances. This aids in filtering the live-captured movies.

2) It can help any video sharing site blur or hide any part with disturbing frames, or erase any video that contains dangerous clips.

3) It may also aid in the creation of online parental control tools, such as browser extensions or plugins that automatically filter out harmful information for children.

Additionally, our technique for identifying objectionable kid-friendly content on YouTube is not influenced by the metadata of videos, which may be readily changed by malevolent uploaders to trick viewers. In order to further enhance the model's performance by better comprehending the global representations of films, we want to integrate the temporal stream utilising optical flow frames with the spatial stream of RGB frames in the future. In order to target the many forms of unsuitable children's material in YouTube videos, we also want to expand the classification labels.

REFERENCES

[1] L. Ceci. YouTube Usage Penetration in the United States 2020, by Age Group. Accessed: Nov. 1, 2021. [Online]. Available: https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/

[2] P. Covington, J. Adams, and E. Sargin, ‘‘Deep neural networks for YouTube recommendations,’’ in Proc. 10th ACM Conf. Recommender Syst., Sep. 2016, pp. 191–198, doi: 10.1145/2959100.2959190.

[3] M. M. Neumann and C. Herodotou, ‘‘Evaluating YouTube videos for young children,’’ Educ. Inf. Technol., vol. 25, no. 5, pp. 4459–4475, Sep. 2020, doi: 10.1007/s10639-020-10183-7.

[4] J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott, Social Media, Television and Children. Sheffield, U.K.: Univ. Sheffield, 2019. [Online]. Available: https://www.stac-study.org/downloads/ STAC_Full_Report.pdf

[5] L. Ceci. YouTube—Statistics & Facts. Accessed: Sep. 01, 2021. [Online]. Available: https://www.statista.com/topics/2019/youtube/

[6] M. M. Neumann and C. Herodotou, ‘‘Young children and YouTube: A global phenomenon,’’ Childhood Educ., vol. 96, no. 4, pp. 72–77, Jul. 2020, doi: 10.1080/00094056.2020.1796459.

[7] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, Risks and Safety on the Internet: The Perspective of European

Children: Full Findings and Policy Implications From the EU Kids Online Survey of 9-16 Year Olds and Their Parents in 25 Countries. London, U.K.: EU Kids Online, 2011. [Online]. Available: http://eprints.lse.ac.U.K./id/eprint/33731

[8] B. J. Bushman and L. R. Huesmann, ‘‘Short-term and long-term effects of violent media on aggression in children and adults,’’ Arch. Pediatrics Adolescent Med., vol. 160, no. 4, pp. 348–352, 2006, doi: 10.1001/archpedi.160.4.348.

[9] S. Maheshwari. (2017). On YouTube Kids, Startling Videos Slip Past Filters. The New York Times. [Online]. Available: https://www.nytimes.com/ 2017/11/04/business/media/youtube-kids-paw-patrol.html

[10] C. Hou, X. Wu, and G. Wang, ‘‘End-to-end bloody video recognition by audio-visual feature fusion,’’ in Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV), 2018, pp. 501–510, doi: 10.1007/978-3-030-03398- 9_43.

[11] A. Ali and N. Senan, ‘‘Violence video classification performance using deep neural networks,’’ in Proc. Int. Conf. Soft Comput. Data Mining, 2018, pp. 225–233, doi: 10.1007/978-3-319-72550-5_22.

[12] H.-E. Lee, T. Ermakova, V. Ververis, and B. Fabian, ‘‘Detecting child sexual abuse material: A comprehensive survey,’’ Forensic Sci. Int., Digit. Invest., vol. 34, Sep. 2020, Art. no. 301022, doi: 10.1016/j.fsidi. 2020.301022.

[13] R. Brandom. (2017). Inside Elsagate, The Conspiracy Fueled War on Creepy YouTube Kids Videos. [Online]. Available: https://www.theverge.com/2017/12/8/16751206/elsagate-youtube-kids-creepy-conspiracytheory

[14] Reddit. What is ElsaGate? Accessed: Dec. 14, 2020. [Online]. Available: https://www.reddit.com/r/ElsaGate/comments/6o6baf/

[15] B. Burroughs, ‘‘YouTube kids: The app economy and mobile parenting,’’ Soc. media+ Soc., vol. 3, May 2017, Art. no. 2056305117707189, doi: 10.1177/2056305117707189.