International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Novel Deep Learning Technique for Detection of Violent Content in Videos

Asmita Poojari [1], Pallavi K. N. *[2], McEnroe Ryan Dsilva [3], Jagadevi N. Kalshetty[4]

Submitted: 09/05/2023 **Revised**: 18/07/2023 **Accepted**: 07/08/2023

Abstract: With a data of 2.5 quintillion bytes worth of data that is being generated daily, regulation of media uploaded on the social media sites such as Facebook, Instagram, and Reddit is a challenge. In addition to social media platforms, there are also private messaging platforms like WhatsApp and Microsoft Teams which are used by private companies for information exchange, team collaborations, and team conversations which must be content regulated. Content moderation is performed by people (moderators) who have to manually classify content into safe and not safe for work. The exposure of human content moderators to harmful and violent-content over the internet makes moderation less desirable. In this study, we suggest and create a Machine Learning Model that can recognise violent video content and classify them as violent and non-violent. Audio and video are the two parameters we use as inputs. The input video along with the audio is initially processed, if the audio is classified as violent, then the video is marked and classified as violent video. The associated video is put into a video classifier where it is further classed as violent or non-violent if the audio classifier deems the audio to be non-violent. Performance indicators like precision, accuracy, sensitivity, and specificity are used to demonstrate the performance of the suggested model.

Keywords: *Content moderation, video classifier, specificity, violent video*

1. Introduction

The introduction of social media into the lives of humans has changed our way of living forever. Social media has affected majorly in the decision making of humans. Most of our lives have been controlled by social media in one way or the other. Social media has an impact on every aspect of our lives, including politics, current events, sports, movies, and even our personal life.

The rise of mobile connectivity and smartphone technology has had a significant impact on the social media landscape. This has resulted in a widespread increase in social media usage. The corona virus pandemic has further contributed to its rise. Reports show that most of the social media websites have seen over 70% gain in popularity during the pandemic. Businesses all around the world are evolving. Applications like Google Meet, Microsoft Teams, and others have grown to be crucial components of their technology stack. With potentially infinite scaling the amount of data and users is huge. Various types of content can be uploaded by users to the social media platform. Violent content must be moderated since some of the content posted to the website may not comply with community standards.

The process of content moderation involves checking for violent material and other factors in content that social media users create and upload to the relevant social media sites. Human moderation is the term used for the practice of manually reviewing each video or image. This method is suitable when the quantity of data is low but due to the rise in popularity of the applications, it becomes a major problem. Large corporations so frequently employ combination of human and AI powered moderation procedures, only using human moderators when the AI is unable to categorise the content correctly.

This paper explains the creation of content moderation software which would be able to flag a video as violent or non-violent by considering both audio and visual features. To prove the performance of the proposed model performance metrics such as precision, accuracy, sensitivity and specificity are used.

2. Related Work

The major reasons for installing video surveillance systems at schools, prisons, hospitals and other places are to alert the officials about the violence being carried on in that place. If human moderation is used in these places, they will be exhausted by the enormous amount of video clippings and the surveillance may not be accurate, hence there arises a

[1] *NITTE (Deemed to be University), Dept. of Computer Science and Engineering, NMAM Institute of Technology, Nitte - 574110, Karnataka, India; asmitapoojari@nitte.edu.in*
ORCID ID : 0000-0003-2994-1958
[2] *NITTE (Deemed to be University), Dept. of Computer Science and Engineering, NMAM Institute of Technology, Nitte - 574110, Karnataka, India; pallavi@nitte.edu.in*
ORCID ID : 0000-0002-3514-2565
[3]*Cloud Associate, Niveus Solutions, Udupi-576101, Karnataka, India; mcenroeryandsilva23@gmail.com*
[4] *Dept. of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bengaluru-560064, Karnataka , India; jagadevi.n.kalshetty@nmit.ac.in*
ORCID ID : 0000-0001-6466-4197

** Corresponding Author Email: pallavi@nitte.edu.in*

need for automatic moderation. Yakaiah Potharaju et.al (2019)[1] proposed violence detection framework which is accomplished under two stages, i.e, feature extraction and classification. Obtained feature set was processed for classification with a supervised learning algorithm. The proposed approach obtained an improvement in the accuracy with 2.2421% and 1.8919%.

In Theodoros et.al (2006)[2] a multimodal approach is used and discussed for the characterization of media contents based on the violence present. Feature extraction is performed and the audio is divided into segments which are further processed frame by frame. The feature sequences are used to produce the eight statistics. These data are used by the classifier as single-feature values for each audio clip.The proposed model showed good results with 85.5 percent of the audio inputs classified correctly on an average.

When a person witnesses violent and bloody scenes in videos, it causes spiritual, emotional and visual influences on the health of small children. This requirement emphasizes the need for video classification and management to protect minor children. Thus H. Wang et.al. (2017)[3] used a multi-modal feature composed of motion, and auditory components, each of which expresses a separate collection of low- and high- level properties. When compared to individual manual features, it was discovered that the high-level features computed from the convolutional neural network have a comprehensive representation.

The authors in Accattoli et.al (2020) [4] proposed a model using the C3D algorithm and trained the model accordingly in order to achieve better results compared to existing algorithms. Suggested violence detection system consists of C3D model customization and a Linear Support Vector Machine classifier which divides the video into 16 frames and provides it as an input to the neural network classifier which utilizes these features to process the frames for violence in contents. The results show that our method achieves excellent accuracy on the datasets for hockey fights and crowd violence, with an accuracy of 98.6% compared to 98.51% for crowd fighting.

In Vosta et.al (2022) [5], The authors intended to process surveillance camera feed and help try to monitorthe feed by reducing the time, money and effort requiredby designing an automated model. To find anomalies in the video collection, a recurrent network (RNN) dubbed Convolutional LSTM (ConvLSTM) is employed. The ResNet50 output(i.e., ConvLSTM) is then obtained by RNN. The suggested solution performs better on the "UCF-Crime" dataset than the other available alternatives.

When it comes to regulating sensitive media content, violence identification is an essential application for video analysis. When used in combination with video surveillance systems, it can be a useful tool forsafeguarding people from being exposed to undesired media from a number of sources. The idea is to developa model to help monitor the same. Bruno Peixoto, et.al.(2020) suggested method decomposes the detection of violence into 'k' more objective sub-concepts that convey the experience of violence. Neural network is trained to first analyze its visual characteristics and subsequently its audio features. The two traits are then combined to identify the violence. Violence detection was a difficult process and the accuracy rate was fairly low.

3. Proposed System Design and Implementation

The proliferation of violent content in videos has become a growing concern in recent years, as it can have harmful effects on individuals and society as a whole. Manual detection of violent content in videos is time-consuming and can be subjective. Therefore, it is necessary to create an automated system that can reliably identify violent content in videos using deep learning algorithms. The ideology is to develop a deep learning model that effectively distinguishes between the violent and the non-violent content in videos.

A detailed analysis and implementation of 2 Deep Learning models is presented, which can be divided into 2 components – Sound Processing and Video Processing. Main objective was performance evaluation of these models in accurately detecting and localizing objects in input data.

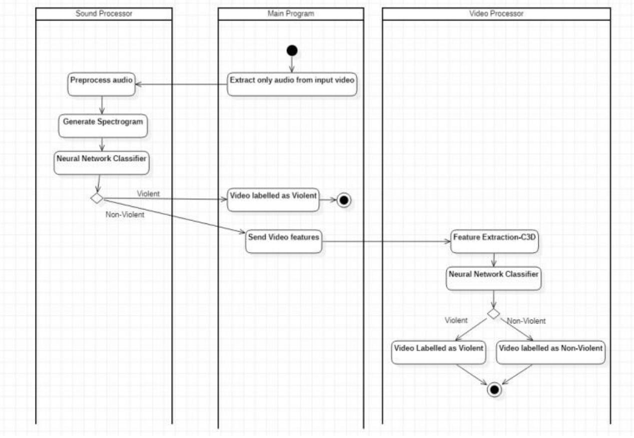

Sound Processor
Main Program
Video Processor
Extract only audio from input video
Preprocess audio
Generate Spectrogram
Neural Network Classifier
Violent
Non-Violent
Video labelled as Violent
Send Video features
Feature Extraction-C3D
Neural Network Classifier
Violent
Non-Violent
Video Labelled as Violent
Video labelled as Non-Violent

Fig. 1. System design of the Proposed method

3.1. Data Collection

Following are the datasets used to train and test the model:

Violent Crowd Dataset: This dataset contains 246 YouTube video clips of various circumstances and scenarios. The dataset starts with five different sample sets of videos. Every set is divided into two categories: non- violent and violent.

Violence in movies dataset: It includes 200 video sample clips, including person-to-person fight videos obtained from action movies and videos classified as non-fight retrieved from publicly available action recognition datasets.

Real life violence situations: This dataset includes more than 2000 films, 1000 of which are violent and 1000 of which are not. which will be primarily utilized to train the model to develop and help learn to distinguish the content put forward later into the right category.

Hockey Fights Dataset: The dataset consists of a variety of hockey and ice hockey videos from various sporting competitions proposed by Simone et.al[4]

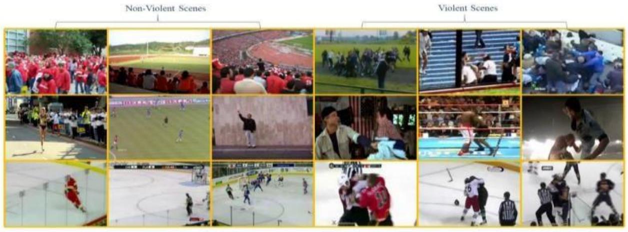


Fig. 2. Data Collection of Hockey Fights dataset

3.2. Audio Processor

Convert: Convert the video format file (MP4, AVI) to a lossless compression format, WAV, which is best suited for audio analysis and processing.

Load audio and video: Reads the input audio waveform (plot of amplitude and sample rate) and converts it into a tensor format.

Pre-process: Every WAV file will be read and its corresponding waveform will be generated which will be fed to the model as an input to generate a spectrogram.

Spectrogram: This returns the spectrogram of the input waveform, which is a tensor waveform. Spectrogram helps visualize amplitude as the function of time and frequency. Time is represented by the x-axis, frequency by the y-axis, and amplitude by the colour. Spectrogram images are utilized to classify videos as "Violent" or "Non-Violent" content. The spectrogram is reshaped from (1900, 129) to (1900, 129, 1).

Audio classifier: The classifier will process the output generated from the model and will detect whether the audio is violent or not.

3.3. Video Processor

Chunks: The number of frames and FPS (frames-per-second) are determined by reading each video in the dataset. Consider 16 frames to be one chunk. As a result, a video's chunk size is calculated by dividing the total number of frames by 16.

Pre-process video: Each video is stored as a NumPy memory map, and every frame in each chunk is resized to a 112x112 shape.

Create and load C3D model: Create the C3D neural network. In order to increase the C3D model's correctness, we loaded the weights pre-trained on the Sports-1M dataset.

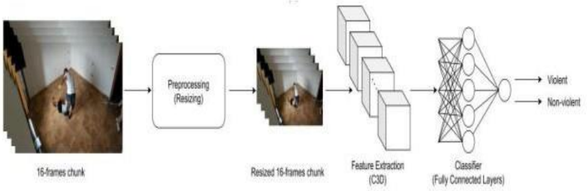


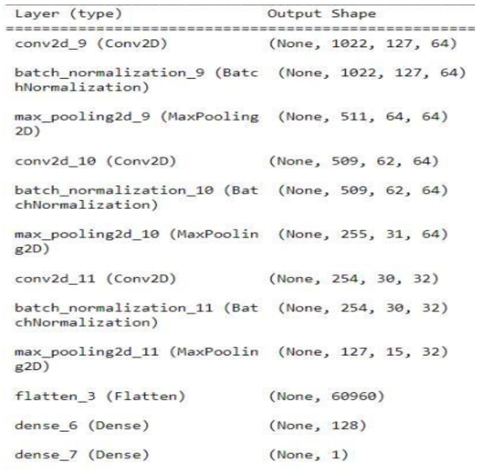
Fig. 3. Video processing model

3.4. Running Model End-to-End

1. We implemented the C3D (Convolutional 3D) neural network architecture for video classification consisting of a deep neural network which consists of one input layer, three convolutional layers, two dense layers, and one output layer. To have reduction in over fitting and to improve the robustness, added Dropout and Batch- Normalization layers between each layer.
2. We used ReLU (Rectified Linear Unit) activation is used after each convolutional layer, and dropout with a rate of 0.5 after the first fully connected layer

to reduce overfitting. The fully connected model is trained in a stratified shuffle split cross-validation scheme. In each split, 70% of the dataset is utilized for training, 10% for validation and 20% for testing.

The metrics presented include accuracy, recall (sensitivity), precision, specificity.



```
Layer (type)                 Output Shape
==================================================
conv2d_9 (Conv2D)            (None, 1022, 127, 64)

batch_normalization_9 (Batc  (None, 1022, 127, 64)
hNormalization)

max_pooling2d_9 (MaxPooling  (None, 511, 64, 64)
2D)

conv2d_10 (Conv2D)           (None, 509, 62, 64)

batch_normalization_10 (Bat  (None, 509, 62, 64)
chNormalization)

max_pooling2d_10 (MaxPoolin  (None, 255, 31, 64)
g2D)

conv2d_11 (Conv2D)           (None, 254, 30, 32)

batch_normalization_11 (Bat  (None, 254, 30, 32)
chNormalization)

max_pooling2d_11 (MaxPoolin  (None, 127, 15, 32)
g2D)

flatten_3 (Flatten)          (None, 60960)

dense_6 (Dense)              (None, 128)

dense_7 (Dense)              (None, 1)
```

Fig. 4. Model Overview

4. Solutions and Results

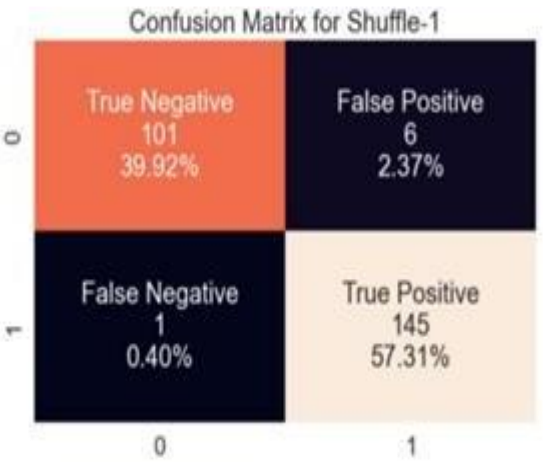
As mentioned before, Stratified Shuffle Cross Validation is mainly used for evaluation of the performance of the proposed approach. During preprocessing, the audio, video features and their labels are stored as NumPy memory-maps. A memory-map is exactly like a HashMap with unique keys and values. 70% of the dataset is utilised for training the model, 20% is used for testing, and 10% is used for validation. In Stratified Shuffle Cross validation, the indices or keys are chosen randomly for training, testing and validation. The process is repeated multiple times and the average of the individual results is presented as the final result. Accuracy, recall (sensitivity), precision, specificity, and the f1-score are among the measures that are reported.

4.1. For Violence Flows Dataset

This dataset was proposed by Dimitris Kosmopoulos, et.al. [2]. There are 246 YouTube video clips in it that show a range of topics and circumstances. Each video clip has a size of 320x240 pixels and a runtime of 50 to 150 frames. Each video is divided into 16-frame segments, and the full dataset is saved as a single NumPy memory-map. It so receives 1265 such chunks. To separate the available chunks into training, testing, and validation datasets, it uses stratified shuffle cross validation. 885 randomly chosen pieces make up the training dataset, 127 chunks make up the validation dataset, and 253 chunks make up the test dataset. Three times through this process, the findings are shown in Table 1 as a consequence.

Table 1. Performance Analysis for Dataset 1

Metric	Shuffle-1	Shuffle-2	Shuffle-3	Average
Accuracy	97.233 %	98.419 %	92.885 %	96.179 %
Precision	97.5 %	100.0 %	94.5 %	97.33 %
Recall	99.315 %	97.314 %	100 %	98.876 %
F1 - Score	97.643 %	98.639 %	94.193 %	96.825 %
Specificity	94.392 %	100 %	83.177 %	92.523 %



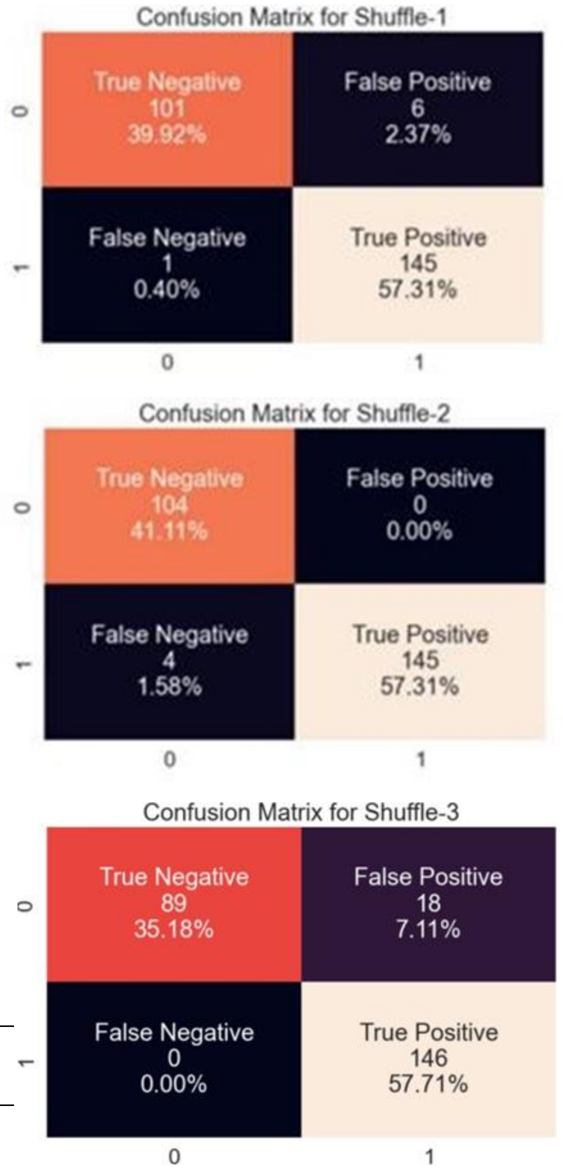


Fig. 5. Confusion matrix for dataset 1

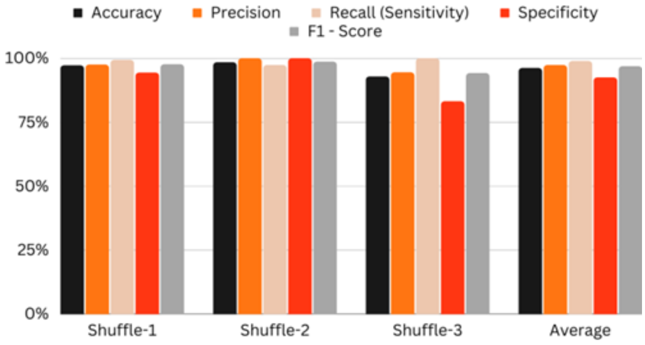


Fig. 6. Graphical representation of performance metrics for dataset 1

On comparison with the original approach proposed introduced in Dimitris Kosmopoulos, et.al. [2] It has been shown that the suggested strategy produces more accurate findings. The proposed approach has a maximum accuracy of 98.419%, while the approach indicated in Dimitris Kosmopoulos, et al. [2] has a maximum accuracy of 96.4% (Table 1). Additionally, the new approach is less prone to over-fitting becausethe proposed approach uses 50 movies for testing and 173 videos for training.

4.2. Violence in Movies Dataset

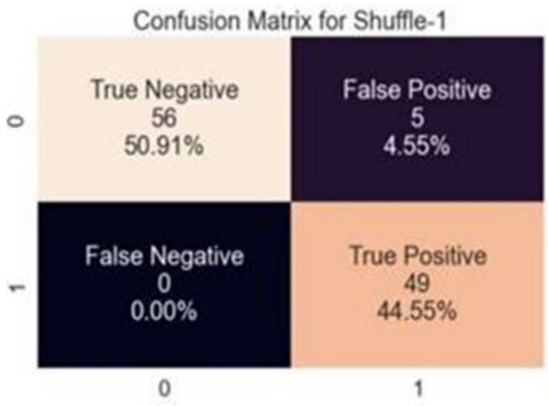
The dataset contains 200 video clips consisting of person-to person fight videos obtained from action movies and videos classified as non-fight retrieved from publicly available action recognition datasets. The videos typically feature 50 frames per second. Later it is segmented every video into chunks of size - 16 frames. Finally, obtains 550 such chunks which are divided into ‘Violent’ and ‘Non-Violent’ categories.

The Stratified Shuffle Cross-Validation is used topartition the available chunks into training, testing, and validation datasets. 385 random chunks are chosen as training samples, 110 chunks are used as testing samples and 55 chunks are used for validation. This process is repeated 3 times and the results are noted as in Table 2.

Table 2. Performance Analysis for Dataset 2

Metric	Shuffle 1	Shuffle 2	Shuffle 3	Average
Accuracy	95.45 %	98.18 %	99.09 %	97.57 %
Precision	95.5 %	98.5 %	99 %	97.67 %
Recall	100 %	95.91 %	100 %	98.68 %
F1 - Score	91.80 %	100 %	98.36 %	96.72 %
Specificity	95.14 %	97.91 %	98.98 %	97.35 %

Since there are no audio streams in the video clips in this dataset either, only the "Video Processor" is in use. It observed that proposed strategy delivers outcomes that are extremely similar to those of the strategies stated in H. Wang. et al. [3] who achieve 100 % accuracy but our method trains with a relatively small number of parameters. As a result, our strategy requires less time tolearn and is more effective.



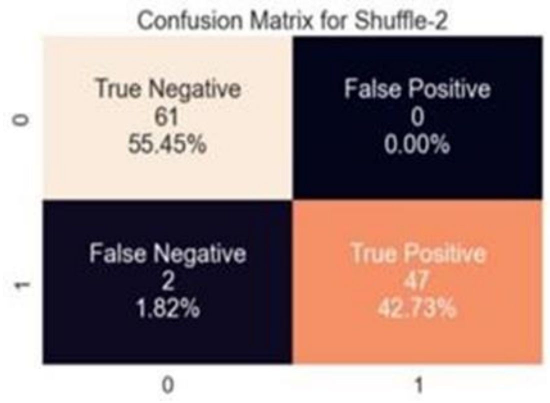


Fig. 7. Confusion matrix for dataset 2

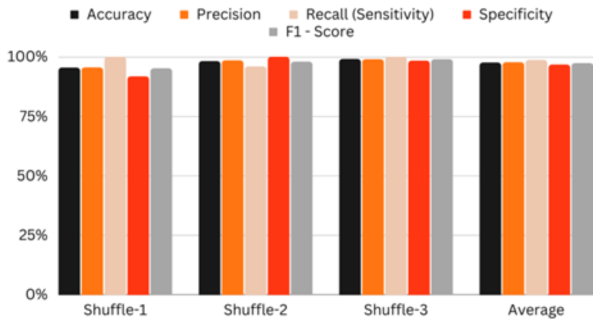


Fig. 8. Graphical representation of performance metrics for dataset 2

5. Conclusion and Future Work

Two deep-learning models for spotting violence in videos are shown in this article. A deep learning model called "Audio Processor" is based on computer vision and uses the spectrogram of audio clips to distinguish between violent and non-violent audio. “Video Processor” is a deep-learning model which uses C3D as a feature extractor and layers of densely connected neurons as the classifier.

Using transfer learning, the weights for the feature extractor are imported from the model trained on the Sports-1M dataset. We have employed 4 benchmark datasets for evaluating our approach. Experimental findings demonstrate that, for both person-to-person fighting and mob violence, our suggested technique outperforms the state-of-the-art methods.

The proposed model also has a significantly lower number of parameters than the models used in most of the approaches.Therefore, the proposed model is very efficient and capable of real-time processing. There were several issues, such as context loss or disorientation in some instances. Ex: The Sound Processor may classify the popping of balloons at gatherings as violent, which led to incorrect findings

Future development might generally consist of the following:

- Implementing a multi-class classifier for recognizing therange of violence present
- Optimizing the approach by using strategies such as sliding window
- Incorporating the model along with social-media applications for content moderation system in further studies.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Yakaiah Potharaju, Manjunathachari Kamsali, Chennakesava Reddy Kesavari. “Classification of Ontological Violence Content Detection through Audio Features and Supervised Learning” International Journal of Intelligent Engineering and Systems, Vol.12, No.3, 2019.

[2] Theodoros Giannakopoulos, Dimitris Kosmopoulos, Andreas Aristidou , S. Theodoridis. “Violence Content Classification Using Audio Features” Advances in Artificial Intelligence, 4th Helenic Conference on AI, SETN 2006, Heraklion, Crete, Greece, May 18-20, 2006, Proceedings

[3] H. Wang, L. Yang, X. Wu and J. He, "A review of bloody violence in video classification," 2017 International Conference on the Frontiers and Advances in Data Science (FADS)

[4] Accattoli, Simone & Sernani, Paolo & Falcionelli, Nicola & Mekuria, Dagmawi & Dragoni, Aldo Franco. (2020). ”Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines”. Applied Artificial Intelligence, February 2020.

[5] Vosta, Soheil, and Kin-Choong Yow. "A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras", Applied Sciences, 2022.

[6] Bruno Peixoto, et.al. “Multimodal Violence Detection in Videos”, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[7] Hospedales, T.; Gong, S.; Xiang, T. Video behaviour mining using a dynamic topic model. Int. J. Comput Vis. 2012, 98, 303–323

[8] Sulman, N.; Sanocki, T.; Goldgof, D.; Kasturi, R. How effective is human video surveillance performance? In Proceedings of the 2008 19th IEEE International Conference on Pattern Recognition, ICPR, Tampa, FL, USA, 8–11 December 2008; pp. 1–3.

[9] Nguyen, T.N.; Meunier, J. Anomaly detection in video sequence with appearance-motion correspondence. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV, Seoul, Korea, 27 October–2 November 2019; pp. 1273–1283

[10] Tian, B.; Morris, B.T.; Tang, M.; Liu, Y.; Yao, Y.; Gou, C.; Shen, D.; Tang, S. Hierarchical and networked vehicle surveillance in its: A survey. IEEE Trans. Intell. Transp. Syst. 2017, 18, 25–48.

[11] Yu, J.; Yow, K.C.; Jeon, M. Joint representation learning of appearance and motion for abnormal event detection. Mach. Vision Appl. 2018, 29, 1157–1170.

[12] Varadarajan, J.; Odobez, J.M. Topic models for scene analysis and abnormality detection. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 1338–1345.

[13] Sodemann, A.A.; Ross, M.P.; Borghetti, B.J. A review of anomaly detection in automated surveillance. IEEE Trans. Syst. Man, Cybern. Part C (Appl. Rev.) 2012, 42, 1257–1272.

[14] Zweng, A.; Kampel, M. Unexpected human behavior recognition in image sequences using multiple features. In Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR, Istanbul, Turkey, 23–26 August 2010; pp. 368–371.

[15] Jodoin, P.M.; Konrad, J.; Saligrama, V. Modeling background activity for behavior subtraction. In Proceedings of the 2008 Second ACM/IEEE International Conference on Distributed Smart Cameras, Trento, Italy, 9–11 September 2008; pp. 1–10.

[16] Dong, Q.; Wu, Y.; Hu, Z. Pointwise motion image (PMI): A novel motion representation and its applications to abnormality detection and behavior recognition. IEEE Trans. Circuits Syst. Video Technol. 2009, 19, 407–416.

[17] Mecocci, A.; Pannozzo, M.; Fumarola, A. Automatic detection of anomalous behavioural events for advanced real-time video surveillance. In Proceedings of the 3rd International Workshop on Scientific Use of Submarine Cables and Related Technologies, Lugano, Switzerland, 31 July 2003; pp. 187–192.

[18] Li, H.P.; Hu, Z.Y.; Wu, Y.H.; Wu, F.C. Behavior

modeling and abnormality detection based on semi-supervised learning method. Ruan Jian Xue Bao (J. Softw.) 2007, 18, 527–537.

[19] Yao, B.; Wang, L.; Zhu, S.C. Learning a scene contextual model for tracking and abnormality detection. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.

[20] Yin, J.; Yang, Q.; Pan, J.J. Sensor-based abnormal human-activity detection. IEEE Trans. Knowl. Data Eng. 2008, 20, 1082–1090.

[21] Benezeth, Y.; Jodoin, P.M.; Saligrama, V.; Rosenberger, C. Abnormal events detection based on spatio-temporal co-occurences. In Proceedings of the 2009 IEEE conference on computer vision and patternrecognition CVPR, Miami, FL, USA, 20–25 June 2009; pp. 2458–2465.

[22] Begum, S. . S. ., Prasanth, K. D. ., Reddy, K. L. ., Kumar, K. S. ., & Nagasree, K. J. . (2023). RDNN for Classification and Prediction of Rock or Mine in Underwater Acoustics. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 98–104. https://doi.org/10.17762/ijritcc.v11i3.6326

[23] Wilson, T., Johnson, M., Gonzalez, L., Rodriguez, L., & Silva, A. Machine Learning Techniques for Engineering Workforce Management. Kuwait Journal of Machine Learning, 1(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/120