International Journal of Technology 14(6) 1228-1236 (2023)
Received June 2023/ Revised September 2023 / Accepted September 2023

International Journal of Technology

http://ijtech.eng.ui.ac.id

A Combination of Light Pre-trained Convolutional Neural Networks and Long Short-Term Memory for Real-Time Violence Detection in Videos

Muhammad Shahril Nizam bin Abdullah[1*], Hezerul Abdul Karim[1], Nouar AlDahoul[2]

[1]*Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia*
[2]*Computer Science, New York University, Abu Dhabi, United Arab Emirates*

Abstract. Machine learning techniques have been used widely to analyze videos that have scenes of violence for censorship or surveillance purposes. Violence detection plays a crucial role in preventing underage and teenage exposure to violent acts and ensuring a safer viewing environment. The automatic identification of violent scenes is significant to classify videos into two classes, including violence and non-violence. The existing violence detection models suffer from several problems, including memory inefficiency and low-speed inference, and thus make them unsuitable to be implemented on embedding systems with limited resources. This article aims to propose a novel combination of light Convolutional Neural Networks (CNN), namely EfficientNet-B0 and Long Short-Term Memory (LSTM). The public dataset which consists of two different datasets, was utilized to train, evaluate, and compare the deep learning models used in this study. The experimental results show the superiority of EfficientNet B0-LSTM, which outperform other models in terms of accuracy (86.38%), F1 score (86.39%), and False Positive Rate (13.53%). Additionally, the proposed model has been deployed to a low-cost embedding device such as Raspberry Pi for real-time violence detection.

Keywords: Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM); Pre-trained model; Violence detection

1. Introduction

Violence detection uses machine learning and computer vision techniques to analyze the videos that are retrieved from surveillance and security cameras or readily available videos, as in the case of censorship of films (Ramzan *et al.,* 2019). Violence detection and censorship are essential acts to prevent children and youths from getting involved in any violent acts, as violent scenes have a significant implication on audience character development and mental state (Hassan, Osman, and Azarian, 2009). The World Health Organization (WHO) has also reported that suicide cases as the second highest determining factor of casualties among people with the age range between 15 to 19 years old across the country (Abbas, 2019).

Artificial intelligence (AI) has been used for various activity recognition, including pornography activities (Hor, *et al.,* 2022) and animal activities (Ong, Connie, and Goh, 2022). In addition, anomaly prediction using machine learning techniques, a subset of AI, was investigated to solve gas transmission (Noorsaman *et al.,* 2023) and electricity

[*]Corresponding author's email: hezerul@mmu,edu,my, Tel.: +60383125499; Fax: +60383183029
doi: 10.14716/ijtech.v14i6.6655

challenges (EL-Hadad, Tan, and Tan, 2022). Consequently, various approaches have been implemented for human activity recognition in videos to identify the normal and abnormal activities (i.e., violent acts) (Sultani, Chen, and Shah, 2018).

This paper presents various classification models including several pre-trained CNNs for feature extraction and LSTM for time series classification, which managed to obtain a higher accuracy compared to the previous experiment made in other research. The models were used to extract frames from videos and classify the sequence of these extracted features into two classes: violence and non-violence. The feature extractors were chosen due to the small number of parameters compared to other available pre-trained models, which helps in the implementation of a system to be deployable into edge devices.

2. Related Work

Most previous works have utilized small-size public datasets, including Movies with 200 videos (Nievas *et al.,* 2011a), Hockey Fights with 1000 videos (Nievas *et al.,* 2011b), and VFD with 246 videos (Hassner, Itcher, and Kliper-Gross, 2012). For instance, The spatiotemporal Encoder was used to detect violent acts based on Bidirectional Convolutional LSTM (BiConvLSTM) architecture (Yu *et al.,* 2019). A combination of CNN and LSTM was trained and evaluated with a public dataset that consists of two different datasets such as RWF-2000 (Cheng, Cai, and RWF, 2021) and RLVS-2000 (Soliman *et al.,* 2019).

The combination of CNN and Convolutional Long Short Term Memory (convLSTM) was investigated using Hockey Fight Dataset, Movies Dataset, and Violent-Flows Crowd Violence Dataset (Sudhakaranand Lanz, 2017). Although good performance was achieved, implementing it on a low-cost Internet of Things (IoT) node presents a challenge. The objective was to detect violence using a memory and computation-efficient, low-cost IoT node (AlDahoul *et al.,* 2021). However, this proposed method utilized a small-sized, arbitrarily defined sequential model during the extraction of features from the dataset in which it only managed to cater less than 80% in terms of accuracy during the prediction over the test batch when it is combined with the LSTM classifier.

3. Materials and Methods

There are two datasets being used in this research, and the description of each is explained in this section. Moreover, the section describes the methodology used in terms of different feature extractors, including pre-trained CNNs as well as classification of extracted features utilizing LSTM. Finally, this section discusses the experimental results and explains further the findings obtained from the said methodology in this work.

3.1. Dataset Overview

Both datasets were obtained from open sources and stored differently by their respective authors. The RWF-2000 dataset (Cheng, Cai, and RWF, 2021) contains videos captured from Closed-Circuit Television (CCTV) cameras, representing diverse real-life situations. This dataset contains 1000 non-violence videos and 1000 violence videos, which were already classified into training and validation datasets. Meanwhile, the RLVS-2000 dataset (Soliman, *et al.,* 2019) comprises videos collected by the author, sourced primarily from online social media platforms like YouTube. In contrast to the RWF-2000 dataset, RLVS-2000 offers a wider variety of video content, including scenes from movies with fighting sequences and pre-acted fighting scenes recorded by groups of individuals to demonstrate fighting scenarios. Additionally, this dataset also contains scenes in different environments and recorded using various devices such as dash cam, mobile phones and so on, and so forth. RLVS-2000 dataset was classified into two (2) different classes, which are

1000 violent videos and 1000 non-violence videos. Ultimately, with the aim to increase the variation as well as to improve the accuracy of detection, this research combined two (2) of the said datasets to compose 4000 videos in total: 3200 videos used for training and validation and 800 used for testing purposes. Figure 1 shows the samples for non-violence videos used in this research, while Figure 2 shows samples of violence videos. The summary of video division into training, validation, and testing is shown in Table 1.

Figure 1 Non-Violence videos thumbnails

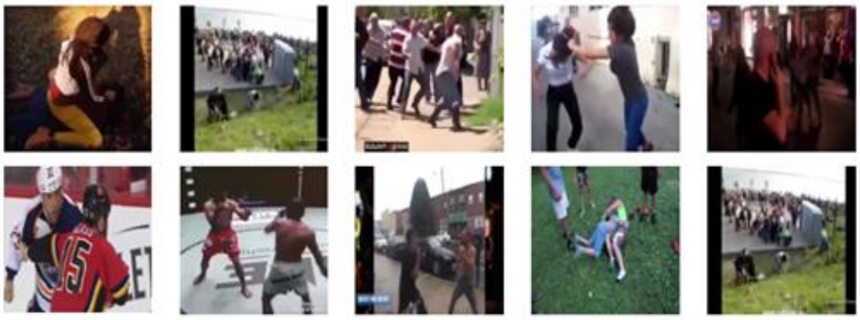


Figure 2 Violence videos thumbnails

Table 1 Summary of videos

Dataset	No. of Videos in Dataset	Use for	Total
RWF-2000	2000	Training	1280
		Validation	320
		Testing	400
RLVS-2000	2000	Training	1280
		Validation	320
		Testing	400
			4000

3.2. Methods

This section demonstrates several pre-trained CNNs that were chosen as the feature extractors due to their small sizes and high accuracy on ImageNet weights. These CNNs include DenseNet-121, EfficientNet B0, MobileNetV2 and ShuffleNet. Additionally, LSTM was added after CNNs to capture a time series of features extracted for violence/non-violence classification purposes.

3.2.1. The solution pipeline

The experiments carried out in this work were divided into several stages, as shown in Figure 3 as follows:

1) Data pre-processing to read video and convert it to sequence of frames for further processing.

2) Feature extraction using various pre-trained CNNs such as DenseNet-121, EfficientNet B0, MobileNetV2 and ShuffleNet-v1 to extract feature vector from each frame.
3) Training in LSTM for classification purposes. The sequence of feature vectors extracted from video frames was applied to LSTM to be classified into violence/non-violence.
4) Model evaluation in terms of accuracy, recall, precision, and F1 score listed in the classification report, with the utilization of the values from the True Positives, True Negatives, False Positives, and False Negatives.

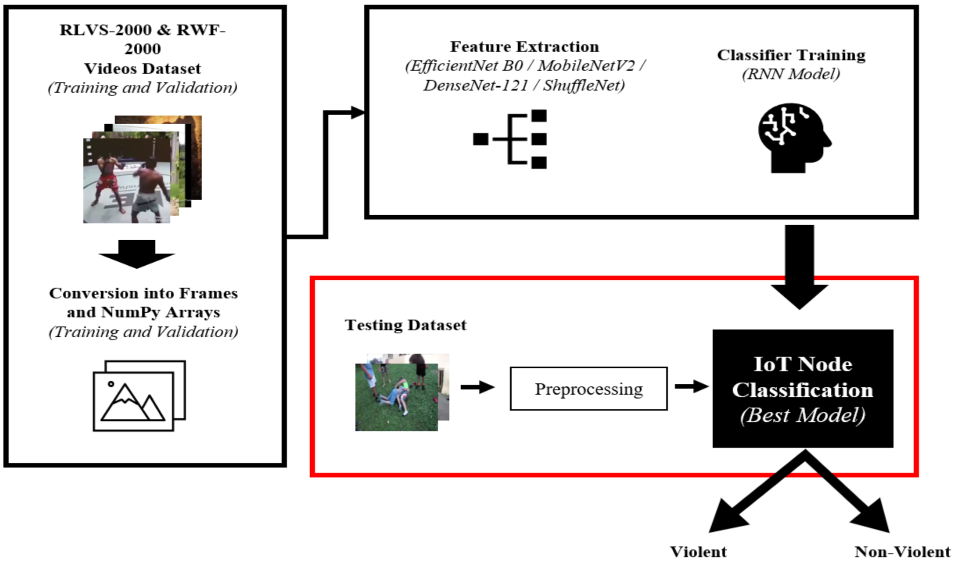


Figure 3 Block diagram for violence detection in videos

4000 videos were converted into Python-accessible format using Python's Open-Source Computer Vision Library or OpenCV. Python can manipulate the OpenCV array structure for characterization if alternative libraries, such as NumPy, are used in a potent mix with it. The vector space and mathematical operations on these features were employed to determine visual patterns and their myriad properties. The videos were pre-processed into the array of NumPy, which was used during the feature extraction and model training.

The experiment was repeated four (4) times with different pre-trained CNNs to determine which combination of CNN and LSTM produced the best result for violence detection in videos. Feature extractors took the converted videos (in terms of NumPy arrays) as the input and passed the input through the available layers according to the layers that were entailed in different pre-trained CNNs and provided different output sizes. These output sizes during the feature extraction determined the number of parameters for our LSTM classifier, as shown in Table 2.

Table 2 Comparison between various feature extractors utilized for violence detection

Feature Extractor	Depth	No. of Parameter	Input Shape	CNN Output Shape
MobileNetV2	105	2,257,984	160x160x3	5x5x1280
EfficientNet B0	132	4,049,571		5x5x1280
DenseNet-121	242	7,037,504		5x5x1024
ShuffleNet	50	969,258		1x1x576

3.2.2. LSTM as the Classifier

The NumPy arrays extracted with the 3-dimensional output shape were reshaped into 2-dimensional NumPy arrays according to their output shape after the feature extraction stage. This reshaped array was used as the input for the LSTM classifier that was created using the Keras Sequential Model with one input layer, stacks of hidden layers, and one output layer. Some of the layers that were used are the LSTM layer, Dense layer, and Activation layer. The number of neurons in each layer was chosen to minimize the model's parameter count while still maintaining good accuracy. Table 3 shows the final LSTM architecture that was used in this work during training, validating, and testing.

Table 3 LSTM Classifier Hyperparameters and Architecture

LSTM Architecture		
Layer	Output Shape	No. of Parameters
LSTM	(None, 10)	1280440
Dense	(None, 512)	5632
ReLU Activation	(None, 512)	0
Dense	(None, 1024)	525312
ReLU Activation	(None, 1024)	0
Dense	(None, 2)	2050
SoftMax Activation	(None, 2)	0

The LSTM Classifier architecture was compiled using the Categorical Cross entropy loss with the “accuracy” metrics monitored. Aside from that, the Adam optimizer was used, with the default learning rate value of 0.01, at 50 epochs with 30 batch sizes.

3.2.3. Hardware implementation for inferencing purpose

Lightweight Pre-Trained CNN and LSTM were combined to perform prediction of violence detection in videos. This would be beneficial to be implemented on devices with limited computability and physical memory. A High-Definition (HD) USB Web Camera was used for capturing the sequence of images in real-time, which acted as an input. Besides that, Raspberry Pi 4 was utilized as the processing unit to perform the prediction over the series of images that were captured in real-time from the webcam using the trained model. An RGB LED was also connected to the Raspberry Pi 4 to act as an actuator to indicate certain colors whenever a violent situation is detected, and vice-versa.

4. Results and Discussion

This section discusses the experimental part of the research, which includes the software and hardware used to set up and implement the methodology. Besides that, this section also focuses on the results of the comparison between numerous feature extractors connected to LSTM for violence and non-violence classification.

4.1. Experimental Setup

The experiments were carried out to train the proposed methods using a laptop that runs on Intel Core i5-7200U at 2.50GHz base frequency and 20GB DDR4 RAM. In other words, these experiments did not involve a Graphical Processing Unit (GPU). The purpose was to assess the suitability of the chosen lightweight pre-trained CNNs for running on edge devices with constrained resources, including limited computation and memory.

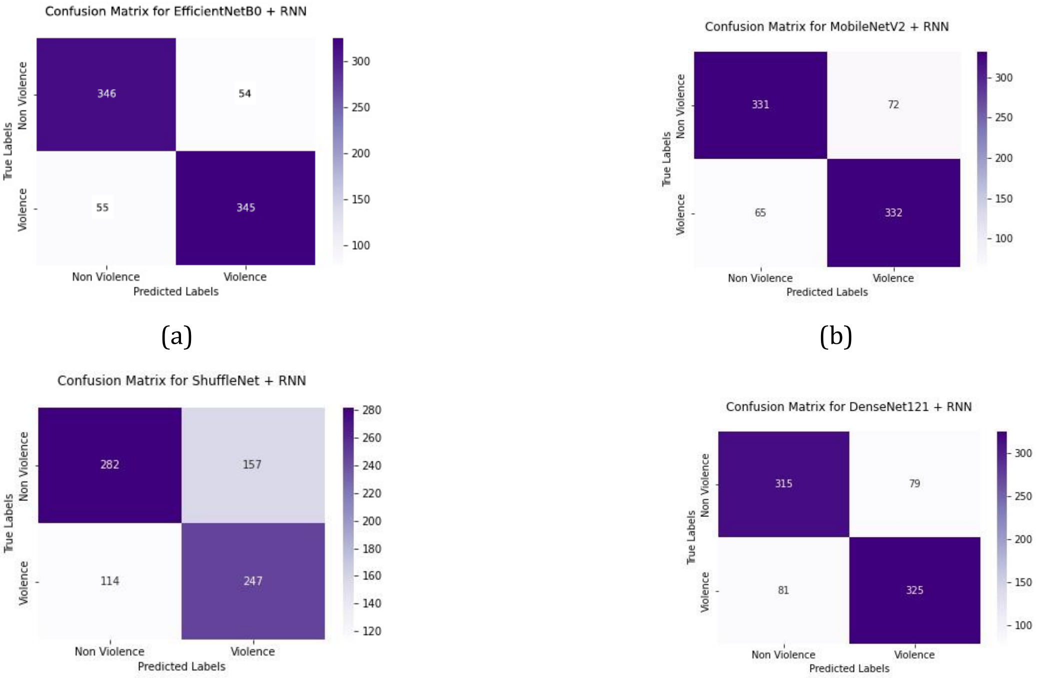
4.2. Experiments and Results

The results of the conducted experiments are presented using the confusion matrix and the classification report. The confusion matrix provides a visual representation of the performance of the classification model, showing the number of true positives, true negatives, false positives, and false negatives for each class that enables the model to be assessed in terms of its ability to correctly classify instances and identify any potential misclassifications. Additionally, classification reports were generated to provide a more detailed analysis of the model's performance. These reports include important metrics such as accuracy, recall, precision, and F1 score, which were calculated for each individual class. By examining these metrics, the insights into the model's accuracy in correctly identifying instances of each class, its ability to recall relevant instances, the precision in correctly classifying instances, and the overall balance between precision and recall as represented by the F1 score were obtained.

4.2.1. Confusion Matrix

Figure 4 shows the results of experiments conducted in this work and presented in terms of confusion matrixes. The experiments have been repeated several times for each feature extractor utilized, and the results presented show the model performance using the testing dataset that contains 800 videos. A confusion matrix is a table of values that displays the number of predictions that the classifier made, in both correct and wrong ways. This helps to evaluate the classifier and calculate several performance metrics for classification, such as accuracy, recall, precision, and F1 score.

Based on Figure 4, the combination of EfficientNetB0 and LSTM classifier outperformed other combinations of CNNs and LSTM. This shows that EfficientNetB0 managed to work well with the LSTM classifier proposed to perform the feature extraction, and classification of videos in terms of violence and non-violence. Furthermore, MobileNetV2 came in the second-ranking regarding the number of correctly predicted videos, with 82.88% of the videos in the test set that were predicted correctly.



(a) (b)

(c) (d)

Figure 4 Confusion Matrix of Combination of Feature Extractor (a) EfficientNet B0 & LSTM Classifier, (b) MobileNetV2 & LSTM Classifier, (c) DenseNet-121 & LSTM Classifier, (d) ShuffleNet & LSTM Classifier

Additionally, DenseNet-121, which was the feature extractor with the highest number of parameters among all the chosen pre-trained CNNs, did not produce the best accuracy, in comparison with EfficientNetB0 and MobileNetV2. There were only 80% of all videos in the testing dataset that were predicted correctly. On the other hand, the confusion matrix of ShuffleNet with the LSTM classifier showed a lot of incorrectly classified videos in the testing dataset. In fact, only 66.13% of videos are classified correctly. This shows the combination of ShuffleNet and LSTM did not perform well compared to the other three feature extractors. Furthermore, the number of False Positive samples and number of False Negative samples shown in the confusion matrix of the ShuffleNet- LSTM classifier were the highest. Hence, it indicated that the ShuffleNet- LSTM classifier did not work well with this dataset.

4.2.2. Classification report

The results were interpreted in the form of a Classification Report that includes several performance metrics such as accuracy, recall, precision, and F1 Score. The calculations of the evaluation metrics in the experiments conducted are based on the value obtained for the True Positives, True Negatives, False Positives, and False Negatives. Table 4 presents the classification report for the compared models based on the testing dataset.

EfficientNetB0-LSTM had the highest accuracy, recall, precision, and F1-score, and the lowest false positive rate, outperforming other combinations of feature extractors and LSTM. In this experiment, EfficientNetB0 achieved an accuracy of 86.38%, indicating the percentage of correctly predicted videos. It also achieved a recall of 86.28%, which represents the rate of true positives captured by the model, and a false positive rate of 13.53%, indicating the rate at which it predicted violence labels for non-violence videos. All the results obtained for EfficientNetB0 lead to the highest F1-score (86.39%), which can be assumed to be the best-performing classification method in this work for detecting violence in the videos. MobileNetV2 and DenseNet-121-LSTM have lower accuracy, recall, precision, and F1-score.

Table 4 Classification Report Results

CNN-LSTM	Accuracy	Recall	Precision	F1-Score	FPR
MobileNetV2 – LSTM	0.8288	0.8359	0.8213	0.8285	0.1782
DenseNet-121 – LSTM	0.8000	0.7955	0.7995	0.7975	0.1955
ShuffleNet – LSTM	0.6613	0.7121	0.6424	0.6754	0.3886
EfficientNet B0 – LSTM (proposed)	0.8638	0.8628	0.8650	0.8639	0.1353

Values, as well as a higher false positive rate. Meanwhile, the combination of ShuffleNet and LSTM classifier showed the lowest accuracy, recall, precision, and F1-score values, and the highest false positive rate compared to the other three combinations. This indicates that this combination is not suitable for detecting violence in videos.

4.2.3. Accuracy Comparison with the Current Implementation

Table 5 shows the comparison of different model performances in terms of accuracy from the proposed work and the currently implemented methodology. Aldahoul *et al.* (2021) have obtained 73.35% accuracy in detecting violent contents using a low-cost IoT node, surpassing the baseline of the pre-trained model used for the feature extraction process. The current proposed work has been made using a few different pre-trained models for feature extraction purposes, and hence, different results have been obtained for comparison of previous methodology over similar dataset utilization. This research has provided a better implementation of violence detection in videos using the low-cost device, whereby the trained EfficientNet B0-LSTM can be made deployable.

Table 5 Comparison of Accuracy With The Current Implementation

Dataset	Model	Accuracy (%)	Author/ Baseline/ Proposed
RWF-2000 + RLVS-2000	CNN Sequential – LSTM	73.35	AlDahoul *et al.* (2021)
	MobileNetV2 – LSTM	82.88	Baseline
	DenseNet-121 – LSTM	80.00	Baseline
	ShuffleNet – LSTM	66.13	Baseline
	EfficientNet B0 – LSTM	86.38	Proposed

5. Conclusions

Lightweight CNN were used to help build a safe environment inside the community by providing methods for violence detection in videos. The current proposed work in this research has made a significant contribution to producing better accuracy for detecting violent contents in videos. The other three combinations of pre-trained models as well as LSTM classifier, outperformed the previous work by at least 6%. EfficientNet B0 has the highest accuracy among the three other combinations, surpassing the previously implemented work by more than 13%. This will lead to better violence detection in videos. For future improvements, these light CNNs can be implemented on edge devices with CUDA-enabled capabilities and include a real-time IoT dashboard to monitor the system. Additionally, the research can be focused on the usage of a light vision transformer to be compared with the light CNNs used in this work for generating violence detection systems with higher speed and accuracy**.**

Acknowledgments

This project is fully funded by Multimedia University.

References

Abbas, A., 2019. Suicide Rate on The Rise, Particularly Among Youth. New Straits Times. Available Online at https://www.nst.com.my/news/nation/2019/09/520301 /suicide-rate-rise-particulaly-among-youth, Accessed on March 13, 2021

AlDahoul, N., Karim, H.A., Datta, R., Gupta, S., Agrawal, K., Albunni, A., 2021. Convolutional Neural Network - Long Short-Term Memory based Internet of Things (IoT) Node for Violence Detection. *In:* 2021 Institute of Electrical and Electronics Engineers (IEEE) International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)

Cheng, M., Cai, K., RWF, M.L., 2021. 2000: An Open Large Scale Video Database for Violence Detection, *In:* 25th International Conference on Pattern Recognition (ICPR), pp. 4183–4190

EL-Hadad, R., Tan, Y.-F., Tan, W.-N., 2022. Anomaly Prediction in Electricity Consumption Using a Combination of Machine Learning Techniques. *International Journal of Technology*, Volume 13(6), pp. 1317–1325

Hassan, M. S., Osman, M. N., Azarian, Z. S., 2009. Effects of Watching Violence Movies on the Attitudes Concerning Aggression Among Middle Schoolboys (13-17 Years Old) at International Schools in Kuala Lumpur, Malaysia. *European Journal of Scientific Research*, Volume 38(1), pp. 141–156

Hassner ,T., Itcher, Y., Kliper-Gross, O., 2012. Violent Flows: Real-time Detection of Violent Crowd Behavior. *In:* 2012 Institute of Electrical and Electronics Engineers (IEEE)

Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6

Hor, S.L., AlDahoul, N., Karim, H.A., Lye, M.H., Mansor, S., Fauzi, M.F.A., Wazir, A.S.B., 2022. Deep Active Learning for Pornography Recognition using ResNet. *International Journal of Technology*, Volume 13(6), pp. 1261–1270

Nievas, E.B., Suarez, O.D., Garcia, G.B., Sukthankar, R., 2011a. Hockey Fight Detection Dataset. In*: Computer Analysis of Images and Patterns,* pp. 332–339

Nievas, E.B., Suarez, O.D., Garcia, G.B., Sukthankar, R., 2011b. Movies Fight Detection Dataset. In: *Computer Analysis of Images and Patterns,* pp. 332–339

Noorsaman, A., Amrializzia, D., Zulfikri, H., Revitasari, R., Isambert, A., 2023. Machine Learning Algorithms for Failure Prediction Model and Operational Reliability of Onshore Gas Transmission Pipelines. *International Journal of Technology*, Volume 14(3), pp. 680–689

Ong, Y.Q., Connie, T., Goh, M.K.O., 2022. A Cow Crossing Detection Alert System. *International Journal of Technology*. Volume 13(6), pp. 1202–1212

Ramzan, M., Abid, A., Khan, H.U., Awan, S.M., Ismail, A., Ahmed, M., Ilyas, M., Mahmood, A., 2019. A Review on State-of-the-art Violence Detection Techniques. *Institute of Electrical and Electronics Engineers (IEEE) Access,* Volume 7, pp. 107560–107575

Soliman, M.M., Kamal, M.H., Nashed, M.A.E.M., Mostafa, Y.M., Chawky, B.S., Khattab, D., 2019. Violence Recognition from Videos using Deep Learning Techniques. *In:* 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 80–85

Sudhakaran, S., Lanz, O., 2017. Learning to Detect Violent Videos using Convolutional Long Short-term Memory, *In:* 14th Institute of Electrical and Electronics Engineers (IEEE) International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6

Sultani, W., Chen, C., Shah, M., 2018. Real-world Anomaly Detection in Surveillance Videos. *In:* Proceedings of the Institute of Electrical and Electronics Engineers (IEEE) Conference on Computer Vision and Pattern Recognition, pp. 6479–6488

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural computation*, Volume 31(7), pp. 1235–1270