Technologique | 2026, Vol. 8, No.1, 1 – 15 | ISSN Online: 3028-1415 | ISSN Print: 3028-1407
DOI: https://doi.org/10.62718/vmca.tech-gjtdsi.8.1.SC-0226-015

OPEN ACCESS

WebVidGuard: A Video Violence Detection System using 3D-CNN and GMVAE

Article History:

Initial submission:	14 February 2026
First decision:	18 February 2026
Revision received:	23 April 2026
Accepted for publication:	30 April 2026
Online release:	07 May 2026

Rengel V. Corpuz[1], ORCID No. 0009-0005-9874-6701
Aleta C. Fabregas[2], DIT, ORCID No. 0000-0003-4983-9985

[1]Master of Science in Information Technology, Polytechnic University of the Philippines, Manila, Philippines
[2]Chief, Center for Computing & Information Sciences Research, Polytechnic University of the Philippines, Manila, Philippines

Abstract

In the digital world, children are increasingly exposed to online video content, including materials that may be violent and harmful. Existing approaches, such as object detection and sequential models, often struggle to capture both spatial and temporal features in video. This limitation frequently results in reduced accuracy, especially when analyzing complex or fast-paced scenes. To address this gap, this study developed WebVidGuard, a real-time video violence detection system in video streams. WebVidGuard combined a 3D Convolutional Neural Network (3D-CNN) with a Gaussian Mixture Variational Autoencoder (GMVAE) to improve detection performance. WebVidGuard was a web-based application that automatically detected and blocked violent video content during playback in real time. The study adopted a mixed-methods approach, combining experiments and development. A dataset of 2,000 video clips, with 1,000 violent and 1,000 non-violent videos, categorized into Punching, Kicking, Head-Hitting, Shooting, and Normal Videos Classes. The dataset was split into 60% training, 20% testing, and 20% validation sets. The system was trained and evaluated using the 3D-CNN and GMVAE models, and performance was measured using accuracy, precision, recall, and F1-score using a confusion matrix analysis across five (5) evaluation runs. The results show that WebVidGuard attained high accuracy and efficiency in detecting violent content. However, performance is affected by the quality and variability of the training data. The study recommends further improvements through dataset expansion, integration of additional modalities, and comparative evaluation with other detection techniques.

Keywords: 3D-Convolutional Neural Network, Gaussian Mixture Variational Autoencoder, Video Violence Detection, Violent Video, Non-violent Video, Object Detection, Action Recognition

INTRODUCTION

Teenagers and school-age youths today are becoming increasingly technologically advanced, often beyond their parents' understanding of modern devices. Exposure to violent media raised serious concerns because of its adverse effects on mental health and behavior (Vinney, 2023). This concern was especially significant among younger audiences, who were more vulnerable to harmful content (Vinney, 2023). Repeated exposure during childhood can lead to long-term aggression and emotional desensitization, even if reduced later (Vinney, 2023). Platforms like YouTube, intended for younger audiences, have also been found to host inappropriate and harmful content (Papadamou et al., 2020). This necessitates robust mechanisms to detect and mitigate exposure to such materials.

Real-time violence detection in videos has gained attention as a viable solution to these challenges. Studies show that 37% of the media targeting children contain scenes of physical or verbal violence. Violence prevalence in movies is at 90%, video games at 68%, TV shows at 60%, and even music videos at 15% (Fitzpatrick, 2018). Guedes and Chavez (2020) demonstrate that dynamic image analysis enables effective violence detection in video content, thus opening ways for systems to proactively detect and block violent content (Guedes & Chavez, 2020). Furthermore, (Hazimah et al., 2020) pointed out how parental control systems would find a significant role in guarding children's

online experience from unwanted access through configurations with wireless networks (Hazimah et al., 2020). This, too, is related to the recent trend for safe digital environments aimed at the younger users. (Fitzpatrick, 2018).

Zhang (2020) analyzed the rapid growth in video-sharing platforms and their impact on viewers. The study explored both the educational potential and harmful effects of short-form video content. The double necessity of using media with constructive purposes while keeping on guard against harmful content lies in this research (Zhang, 2020). Short-form video platforms increase accessibility and allow younger audiences to easily consume user-generated content. These platforms emphasize social interaction, fast content sharing, and simple content creation processes. Videos longer than ten minutes are already classified as long-form (Shutsko, 2020; Zhang, 2020).

Despite advancements, existing methods like YOLO (You Only Look Once) and RNN (Recurrent Neural Networks) shows limitations in analyzing complex video data (Gao, 2023). YOLO splits images into grids and detects bounding boxes and class probabilities for fast, efficient real-time object detection. Yolo efficiently detects objects in images but struggles with temporal patterns and contextual understanding in videos (Guedes & Chavez, 2020). RNNs handle sequential data but face challenges like vanishing gradients and limited long-term dependency learning. Their sequential processing nature also reduces efficiency in real-time video analysis applications (Sabokrou et al., 2017, 2018).

These limitations often result in reduced detection accuracy, particularly in fast-paced or context-dependent scenarios. Furthermore, many existing systems focus either on object detection or sequential modeling but lack an integrated approach that combines both spatial-temporal feature extraction and anomaly detection. This study aims to address these gaps by developing WebVidGuard, a real-time video violence detection system that integrates 3D-CNN and GMVAE. Specifically, the study aims to develop a system capable of detecting violent objects and actions in video streams and evaluate its performance using accuracy, precision, recall, and F1-score.

LITERATURE REVIEW

Effective tools for intelligent video analysis are highly demanded, especially to identify violence in video streams and static objects like blood and cold arms (Peixoto et al., 2019). CNN has been applied extensively in the image domain, greatly enhancing the performance of object detection, action recognition, image classification, etc. The success of CNN in the image domain has inspired research on CNN-based action video recognition (Muhammad et al., 2021; Yao et al., 2019a; Yao et al., 2019b).

The CNN is a feed-forward artificial neural network that is a biologically inspired version of multilayer perceptions (Soliman et al., 2019). It consists of multiple hidden layers, an input layer, and an output layer. The hidden layers may be fully connected, pooling, or convolutional (Tamam et al., 2023). The convolutional layer applies a convolution operation and an additive bias to the input data, passing the result first through an activation function and then onward to the next layer. (Yao et al., 2019a; Yao et al., 2019b).

In computer vision applications such as object detection, CNN-based techniques have dramatically improved (Luo et al., 2019). The real-time object detection outperforms some of the previously suggested models at a 92.7% accuracy rate (Juneja et al., 2021). Some actions of physical harm, such as Kicking, Punching, and Head-hitting could be detected by CNN (Escobanez & Comendador, 2022). Larger and deeper neural networks are expected to consume more memory and processing power. This is particularly troublesome for inference-based systems operating on computer platforms with constrained resources (Hailesellasie & Hasan, 2019). Finding scenes in a video stream with violence is the main challenge of violence detection. The violent acts under consideration may be varied, such as

gunshots, explosions, robberies, assaults, and fighting (Khan et al., 2019).

A novel end-to-end partially supervised deep learning technique enables video anomaly localization and identification using just typical instances (Fan et al., 2020). The method relies on a Gaussian Mixture Variational Autoencoder that can learn feature representations of the normal data through deep learning training (Fan et al., 2020). Song et al., (2019) published a novel approach in violent video detection based on a modified 3D-CNN, which used random sampling methods in dividing nodes to produce input frame sequences. The researchers used three public violent video datasets for their classifier, wherein they used hockey fights, movies, and crowd violence that were individually strategized to suit the length of the clips (Song et al., 2019).

Several studies on video violence detection utilized publicly available datasets such as Hockey Fight, UCF-Crime, and RWF-2000, each with distinct characteristics. The Hockey Fight dataset consists of short clips focused on one-on-one physical alterations, making it suitable for binary classification but limited in diversity (Song et al., 2019). In contrast, UCF-Crimes contains long, untrimmed videos with various real-world anomalies, including violent events, offering greater variability but increased complexity in temporal detection, as reported in prior video anomaly detection studies (Cheng et al., 2020). Meanwhile, RWF-2000 provides a balanced dataset of real-world fight and non-fight scenes with improved diversity in crowd interactions and environmental conditions (Cheng et al., 2020). Compared to these datasets, the study utilizes a curated dataset of 2,000 labeled video clips combining multiple categories of violent actions and objects, allowing for controlled evaluation of both object detection and action recognition performance in a unified framework.

Figure 1 shows a comparison of algorithms in object detection in real time. Among the algorithms are: (1) (QNN) Quantum Neural Network, (2) (CNN) Convolutional Neural Network, and (3) (BNN) Binarized Neural Network. In the study of V.R.S. Mani et al. (2022), the CNN Deep Neural Networks classifier provides 3.458% and 1.600% higher accuracy values compared to other real-time object detection models, demonstrating improved detection performance (Mani et al., 2022). Sangwan & Jain (2019) found that CNN-based approaches perform well in detecting threat-related objects in a complex environment (Sangwan & Jain, 2019). Jain & Vishwakarma (2020) further emphasized the effectiveness of convolutional neural networks in enhancing violence detection accuracy (Jain & Vishwakarma, 2020). These findings collectively support the effectiveness of CNN architectures in object detection and violence recognition.

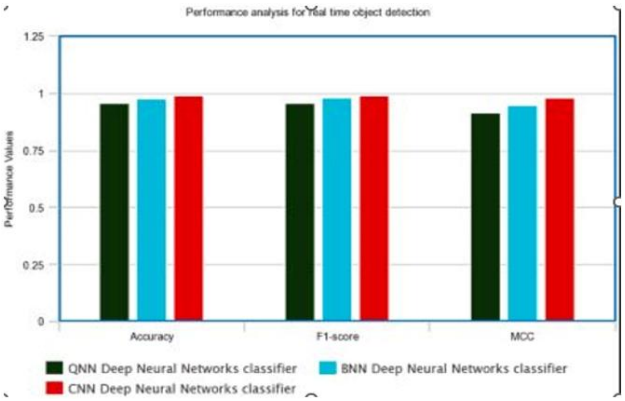

Figure 1
Performance comparison of CNN, QNN, and BNN deep neural networks for real-time object detection using ZYNQ FPGA node. Adapted from Mani et al. (2022).

Convolutional Neural Networks (CNNs) have been widely used in image and video analysis due to their strong capability in feature extraction and classification. Studies by Muhammad et al. (2021) and Yao et al. (2019) demonstrate significant improvements in action recognition using CNN-based architectures (Muhammad et al., 2021; Yao et al., 2019a). Similarly, Escobanez & Comendador (2022) showed that CNNs can effectively detect physical actions such as punching, kicking, and head-hitting (Escobanez & Comendador, 2022). However, unlike these studies, which focus primarily on spatial features, Sabokrou et al. (2017, 2018) emphasized the limitations of CNNs in capturing temporal dependencies, which are

essential in understanding continuous violent actions in videos (Sabokrou et al., 2017, 2018).

Its excellent accuracy and efficiency in picture and object recognition make it a useful feature for media detection, recognition, and feature extraction. To extract information from an input image or sequence, a convolutional neural network (CNN) employs a considerable number of convolutional layers. Each convolutional layer produced a set of convolved feature maps by applying a set of filters, commonly called kernels, to the input picture. CNN has the advantage of learning from raw pixel data instead of requiring human feature engineering techniques. A deep neural network for notable detection accuracy has been developed. CNN has an accuracy rate of 75% on its own and is usually higher when combined with additional methods (Mani et al., 2022). However, unlike standalone CNN approaches, Fan et al. (2020) highlighted that combining deep learning models with clustering techniques such as GMVAE can further enhance performance, particularly in distinguishing complex patterns such as violent and non-violent behaviors (Fan et al., 2020). This suggests that hybrid approaches may provide better generalization compared to single-model architectures.

Since CNNs are excellent at extracting spatial data, they are applied quite often in video violence detection. However, they have major shortcomings, especially concerning the maintenance of temporal linkages that are so crucial for the observation of violent events that may take several frames. The inefficiency of CNNs in processing video sequences when time is of the essence can be attributed to their focus on static image input (Sabokrou et al., 2017, 2018). Furthermore, CNNs tend to overfit and perform poorly on novel scenarios when trained on limited or unbalanced datasets (Fan et al., 2020). Similarly, gradient problems that can vanish or inflate affect the capacity of RNNs, specifically those designed for the analysis of sequential data. It is less efficient in the simulation of important long-term dependencies used for persistent aggressive actions. Hence, the slow processing speed renders RNNs less capable for real-time usage (Sabokrou et al., 2017, 2018).

The latest video analysis developments have come through the application of deep learning technology, such as 3D Convolutional Neural Networks and Gaussian Mixture Variational Autoencoders in the solution to problems of anomaly and violence in videos (Nwokonkwo et al., 2024). Indeed, such models have shown strong potential in making sense of more complex temporal and spatial relationships inherent in video data (Sabokrou et al., 2017, 2018).

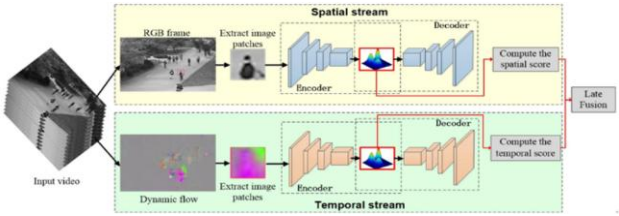

Figure 2
GMVAE Framework (Fan et al., 2020)

The research aimed to overcome the major issues by devising a real-time-based system that conducted violence detection using a combination of 3D-CNN and GMVAE technologies. This system is prepared not only to identify and refrain from accessing violent videos but also to include parental control functionalities, thus offering better usability and responsible media consumption promotion, which would result in child-friendly online environments.

Gaussian Mixture Model (GMM). This statistical model assumes that normal data arises from the mixture of multiple Gaussian distributions. Then, anomalies are identified as points that do not fit well into any of those distributions. Figure 2 shows a Two-stream Gaussian Mixture Fully Convolutional Variational Autoencoder GMFC-VAE is used in learning an anomaly detection model based on dynamic flows and normal samples of RGB images (Fan et al., 2020).

3D Convolutional Neural Networks (3D CNNs). These networks process both spatial and temporal information in video frames for

efficient violence detection (Aremu et al., 2023; Cheng et al., 2020). Figure 3 shows three phases, in which phase 1 is acquiring a video stream from a camera and detecting people. Phase 2 is extracting deep features and then feeding the selected frames to the 3D-CNN model to detect violent activity. Phase 3 is when violent activity is detected; it would take immediate action to report to authorities before any injury or disaster occurs (Ullah et al., 2019).

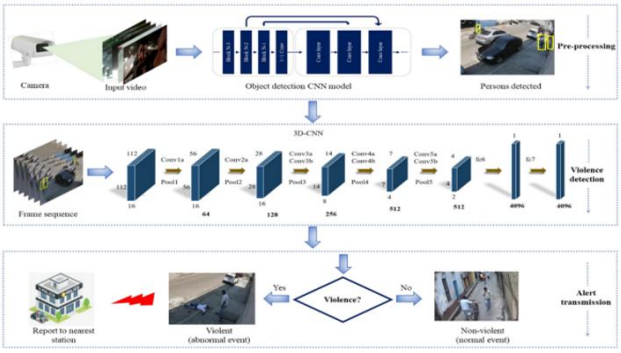

Figure 3
3D CNN model Framework (Ullah et al., 2019)

Although optical flow algorithms and variational autoencoders (VAEs) have much potential, they have disadvantages in video violence identification. VAEs are constructed to reconstruct input data with a compressed latent space, which results in false negatives since they cannot grasp the subtleties of complex violent scenarios. Their capacity to differentiate between violent and non-violent actions is further degraded due to the trade-off between reconstruction fidelity and latent space regularization (Fan et al., 2020; Park et al., 2024).

Optical flow algorithms have attempted to incorporate motion analysis with RGB data to achieve higher accuracy-even those that are based on Conv3D-based frameworks-suffer from unwanted effects of camera jitter and light changes (Park et al., 2024). Among the famous object identification methods is You Only Look Once (YOLO), well known for fast and very effective real-time detection. This method splits photos into grids and concurrently estimates the bounding boxes together with class probabilities. However, the temporal dynamics and complex contextual clues required to identify violent events in movies are hard to handle by YOLO because it is optimized mainly for object identification (Guedes & Chavez, 2020).

In summary, existing studies demonstrate that deep learning techniques such as CNN, RNN, and 3D-CNN have greatly enhanced the efficiency of video violence detection systems. However, limitations remain, particularly in effectively capturing both spatial and temporal features, handling complex real-world scenarios, and reducing false-positive detections. Additionally, most existing systems focus primarily on detection accuracy without integrating real-time deployment or user-oriented control mechanisms such as parental content filtering. These gaps demonstrate the need for a more thorough approach that incorporates reliable feature extraction, anomaly detection, and real-time implementation. To address these limitations, the study proposes WebVidGuard, a web-based video violence detection system that integrates 3D-CNN and GMVAE to enhance detection accuracy and provide automated content blocking for safer online viewing.

METHODS

This chapter discussed the research design, data collection, analysis techniques, and system development of WebVidGuard, a real-time video violence detection system using 3D-CNN and GMVAE.

The study employed an experimental and developmental research design, focusing on system development and performance evaluation. The developmental aspect focused on developing, implementing, and evaluating the developed detection system, while the experimental method approach evaluated system performance using metrics such as accuracy, precision, recall, and processing speed.

Data for this research came from the following sources. Primary data was curated from videos

labeled as violent or non-violent, sourced from publicly available repositories such as UCF-Crime, Hockey Fight Dataset, and Kaggle. This study utilized Dataset Labels and Performance Metrics. Dataset Labels were the ground truth labels for video clips used to train and validate the detection system.

For the datasets, one thousand (1,000) non-violent and one thousand (1,000) Violent video, or a total of two thousand (2,000) videos, were used. The detection model accuracy of the system was tested with a labeled video segment dataset. The video data was preprocessed in the first phase to satisfy the input conditions of the 3D-CNN and GMVAE models. Next, the data was labeled, facilitating the identification of objects within the images.

The video dataset was annotated by an expert in video analysis with a background in computer vision and machine learning. The expert was selected based on prior experience in action recognition tasks and familiarity with violent/non-violent behavior in video data. To ensure labeling reliability, a subset of 200 videos was independently annotated by a second expert, and inter-rated reliability agreement was measured using Cohen's Kappa, achieving a value of 0.87, indicating substantial agreement. This process ensured that the dataset labels were valid and consistent for training and evaluating the model.

A Data Quality Check was performed to assess the integrity and accuracy of the labeled data. The validated dataset was split into 60% training, 20% testing, and 20% validation. This followed the standard practice in machine learning research for medium-sized datasets to balance model learning and unbiased evaluation (Goodfellow et al., 2016). The segments were processed within the system, and performance metrics such as accuracy, precision, recall, and F1-score were determined. The baseline models were implemented using the same dataset, preprocessing steps, and evaluation metrics to ensure a fair comparison.

To analyze the evaluation of system performance, the researcher employed quantitative analysis to statistically treat the gathered data. Accuracy, precision, recall, and F1 score in determining detection performance were used.

Accuracy measured the proportion of correctly predicted violent and non-violent videos (both positive and negative) out of the total predictions.

$$\text{Accuracy} = \frac{True\ Positives\ (TP)\ +\ True\ Negatives\ (TN)}{Total\ Predictions\ (TP\ +\ TN\ +\ FP\ +\ FN)}$$

Precision measured the proportion of true positive violent and non-violent videos predictions out of all positive predictions. It reflected how accurate the model was when predicting violent videos.

$$\text{Precision} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) +\ False\ Positives\ (FP)}$$

Recall (or sensitivity) measured the proportion of true positives, violent and non-violent videos, out of all actual positives. It indicated how well the model identified violent videos.

$$\text{Recall} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) +\ False\ Negatives\ (FN)}$$

The F1 Score was a balanced metric that considered both false positives and false negatives. It was calculated as the harmonic mean of Precision and Recall. When the data was unbalanced, it was very helpful (when there were more violent videos than non-violent videos).

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision +\ Recall}$$

A confusion matrix compared true and predicted values to evaluate a model's performance. It demonstrated the model's ability to predict the true value of each sample.

To reduce the likelihood of viewing violent content, the WebVidGuard was developed to identify violent content in videos. WebVidGuard used 3D-CNN and GMVAE for violent action detection, integrating spatial-temporal feature extraction with unsupervised clustering for robust performance, as shown in Figure 4. Videos that contained violence were restricted using 3D-CNN for feature extraction, object detection, and action recognition. GMVAE was utilized for data clustering in the analysis of what constitutes violence to increase its efficiency. This was an effective way for parents to keep an eye on their kids' internet safety.

	Predicted class 0	Predicted class 1
True class 0	TP	FP
True class 1	FN	TN

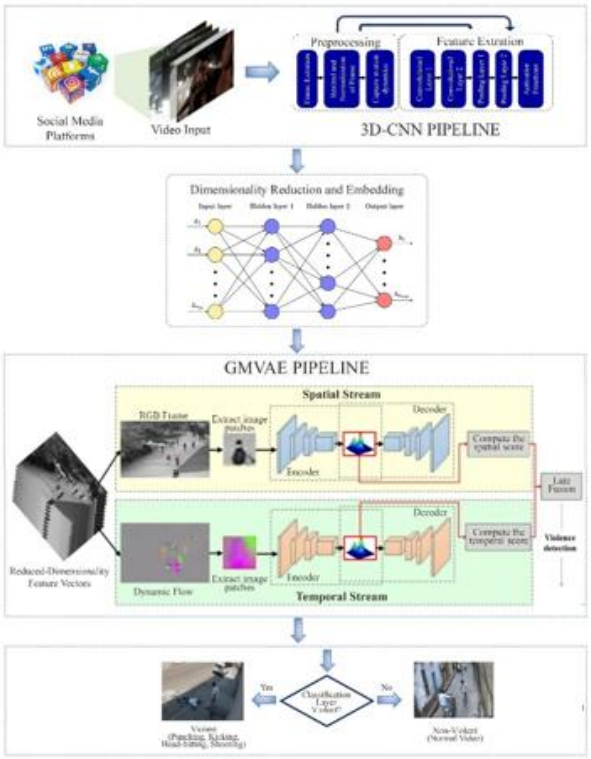

Figure 4
System Architecture of the WebVidGuard

The 3D-CNN model consisted of 5 convolutional layers with kernel sizes of 3x3x3, each followed by ReLU activation and max-pooling layers. The network was trained with a batch size of 16, a learning rate of 0.001, using the Adam optimizer for 50 epochs. The GMVAE model included an encoder with two fully connected layers (512 and 256 neurons) and a latent space of 128 dimensions, with a decoder mirroring the encoder structure. Hyperparameters were selected based on preliminary experiments and prior works in video anomaly detection (Fan et al., 2020). These specifications ensured reproducibility of the model training and evaluation process.

To establish the added value of the combined 3D-CNN and GMVAE architecture, two baseline models were also trained and evaluated on the same dataset. The standalone 3D-CNN model used the same 3D-CNN architecture described for the combined model, but without the GMVAE component. It focused solely on spatial-temporal feature extraction for action recognition and object detection. The standalone GMVAE model employed only clustering and anomaly detection without the spatial-temporal features provided by 3D-CNN. All models were trained and evaluated using the same dataset split and identical performance metrics. This allowed for a controlled comparison to quantify the improvement provided by integrating the models in the system.

The model used in Video Violence Detection was developed using the Python Programming language in Visual Studio Code. For image analysis and machine learning, the researcher used TensorFlow. TensorFlow was a versatile deep learning framework that supported both CNNs and other machine learning algorithms. TensorFlow library was used for building the 3D-Convolutional Neural Network (3D-CNN) and Gaussian Mixture Variational Autoencoder (GMVAE), which was used to detect violence in videos. Images and clips were collected from Kaggle, UCF-Crime, and Hockey Fights, containing violent and non-violent scenes, and were used as training data for the WebVidGuard.

The confidence level of the model referred to the probability associated with the predicted action class. In this study, confidence was derived from the output of the classification

layer of the GMVAE, where class logits were transformed into probabilities using the SoftMax function. The highest probability value corresponded to the predicted class and represented the model's confidence level in identifying a specific action, such as punching, kicking, head hitting, shooting, or non-violent behavior.

The WebVidGuard system was designed as a web-based application that interfaces with a browser environment. It operated as a content filtering mechanism by analyzing video streams during playback. When a video was accessed, frames were extracted and processed by the 3D-CNN + GMVAE model in the backend. If violent content was detected based on predefined thresholds, the system restricted playback by blocking or interrupting the video stream. This allowed WebVidGuard to function as a parental control tool that can be integrated into browser-based platforms for real-time monitoring and filtering of online video content.

Ethical Considerations. A real-time violent detection system has developed a complex ethical landscape that required careful consideration and proactive measures for mitigating risks. The concern of data privacy was addressed, algorithms were ensured to be free from bias, and misuse was prevented so that individual freedoms were respected, and the technology was used responsibly and ethically. The system prevented or excluded other nonviolent activities from recorded data. All video data used complied with data privacy standards without disclosing any personal or sensitive information. Developers ensured responsible and ethical use of the technology by addressing data privacy concerns, eliminating algorithmic bias, preventing misuse, and respecting individual freedoms.

The study obtained ethical approval from the school's Research and Extension Office with approval code: 2025-136. Although the study did not involve human participants directly, ethical considerations were applied to ensure the responsible use of video data. All videos were sourced from publicly available datasets, and no personal or sensitive information was disclosed. The development and evaluation of the system followed ethical research practices to respect data privacy and avoid misuse.

RESULTS

This chapter discussed the study's findings following the researcher's experimentations. Specifically, it presented the WebVidGuard's effectiveness and accuracy rate in detecting violence in videos.

The results of the researcher's experimentation using 60 video samples per attribute per classification task in WebVidGuard. Object detection for gun and knife used a total of 120 video samples. Action recognition for punching, kicking, shooting, and head-hitting used a total of 240 video samples. Temporal Context used 60 video samples, while Speed Impact Accuracy used 60 video samples. Overall, 480 video samples were evaluated for the WebVidGuard system.

The researcher recorded every output from the system and compared it against expert-verified annotations of violent and non-violent content. If the detected attribute (e.g., object, violent action, temporal sequence, or playback speed effect) matched the expert annotation, the model's output was considered correct; otherwise, it was considered incorrect. The evaluation employed a confusion matrix for Accuracy, Precision, Recall, and F1 Score.

Table 1

Confusion Matrix of WebVidGuard for Object Detection

Object	TP	TN	FP	FN	Total	Accuracy	Precision	Recall	F1 Score
Gun	29	25	1	5	60	90.00%	96.67%	85.29%	90.63%
Knife	28	27	2	3	60	91.67%	93.33%	90.32%	91.80%
Total	57	52	3	8	120	90.83%	95%	87.69%	91.20%

The researcher evaluated the object detection by WebVidGuard on the frames in the test folder, specifically on Gun and Knife objects, without depending on the training data. The researcher evaluated object detection by assessing its classification accuracy as well as its localization accuracy using bounding boxes and IoU values.

For Guns, WebVidGuard performed well: it obtained an accuracy of 90%, a precision of 96.67%, and a 90.63% F1 score. This revealed that it had a very high level of confidence with few false positives. However, it obtained a recall of only 85.29%, which shows that some instances were missing.

In the case of Knives, the result was also satisfactory, with an accuracy, precision, and recall of 91.67%, 93.33%, and 87.69%, respectively. The F1-score stood at 91.20%. Thus, the localization of a knife object was as reliable as the localization of a gun object, apart from a minor drop in the recall that reveals some knife objects were overlooked. Overall, the model reliably detects and localizes the dangerous objects within the video with strong consistency for both types.

During action recognition validation, the researcher focused on four classes: "Shooting", "Punching", "Kicking", and "Head Hitting". The test dataset used similar videos created during training to reflect realistic and practical real-world scenarios.

Table 2

Confusion Matrix of WebVidGuard for Action Recognition

Action Recognition	TP	TN	FP	FN	Total	Accuracy	Precision	Recall	F1 Score
Punching	29	27	1	3	60	93.33%	96.67%	90.63%	93.55%
Kicking	28	27	2	3	60	91.67%	93.33%	90.32%	91.80%
Head hitting	26	25	4	5	60	85.00%	86.67%	83.87%	85.25%
Shooting	28	29	2	1	60	95.00%	93.33%	96.55%	94.74%
Total	111	108	9	12	240	91.25%	92.50%	90.24%	91.36%

Between the four video action classes, WebVidGuard registered its highest performance with Shooting by recording an accuracy of 95%. This was coupled with an F1-score of 94.74%. This revealed the ability of the linkage to identify the violent act with recall, combined with precision. Punching, contrary to expectations, came a close second with an accuracy of 93.33%. This was coupled with a high F1-score of 93.55%. Kicking activity remained consistent with an accuracy of 91.67%. This revealed its ability to identify non-violent acts. Head-hitting has the lowest accuracy of 85.00% among the four violent actions. This may be due to the visual similarity between violent head-hitting movements and normal head motions, as well as occasional occlusion in crowded frames, which can confuse the classifier. This aligns with prior studies highlighting the challenges of detecting subtle or partially obscured violent actions in video streams (Sabokrou et al., 2017, 2018).

Overall, the findings indicate that the WebVidGuard works efficiently to identify violent, as well as non-violent, activities within scenarios that are rich with context and structurally sound. Compared to YOLO-based approaches (Gao, 2023), which primarily focus on object detection, WebVidGuard achieved higher accuracy in recognizing complex violent actions due to its 3D-CNN component capturing temporal features. Unlike RNN-based systems (Sabokrou et al., 2017), which are limited by sequential processing and vanishing gradient, the hybrid 3D-CNN + GMVAE model maintained both high precision and recall across multiple action classes. The researcher also noted the prominence of the influence of frame continuity and volumes on the predictions, particularly when the situations are intricate. More frames were found to increase the accuracy and detection of activities such as physical assault and shooting. These comparisons demonstrate the advantage of integrating spatial-temporal extraction with unsupervised clustering in real-time violence detection.

Table 3

Confusion Matrix of WebVidGuard for Temporal Context

Category	TP	TN	FP	FN	Total	Accuracy	Precision	Recall	F1 Score
Temporal Context	22	31	3	4	60	88.33%	88%	84.62%	86.28%

In terms of temporal context, as shown in Table 3, the WebVidGuard achieves an accuracy of 88.33%, confirming that the combined architecture effectively preserved and interpreted sequential frame information compared to frame-based methods. The model obtained a precision of 88% and an F1 score of 86.28%, suggesting a balanced performance with relatively low false-positive rates. However, the recall was slightly lower at 84.62%, indicating that some instances of temporal patterns were not detected. This result implies that while the model is effective in correctly identifying temporal features, there

is still room for improvement in capturing all relevant sequences of violent actions.

Table 4
Confusion Matrix of WebVidGuard for Speed Impact Accuracy

Category	Normal Speed		Fast-Forward Speed		Accuracy_ normal	Accuracy_fast-forward	Impact Accuracy
	TP	TN	TP	TN			
Speed Impact Accuracy	28	27	26	26	91%	87%	95.60%

For speed impact accuracy, as shown in Table 4, the WebVidGuard achieved an accuracy of 91% under normal playback speed conditions and 87% under fast-forward playback speed, reflecting a 4% decrease in performance. This corresponds to an Impact Metric Accuracy Percentage of 95.60%, indicating that the model retains most of its detection capability despite changes in playback speed. These results suggest that the WebVidGuard demonstrates robustness in handling variations in video speed, maintaining reliable performance in dynamic viewing conditions.

Table 5
Confidence Level Distribution

Category	Class	Mean Confidence	Std Dev
Object Detection	Gun	82.54%	19.92%
Object Detection	Knife	79.51%	16.79%
Action Recognition	Punching	92.89%	12.65%
Action Recognition	Kicking	82.73%	30.98%
Action Recognition	Head hitting	71.49%	23.90%
Action Recognition	Shooting	93.11%	10.63%

Table 5 presents the class-wise confidence levels computed from the model's prediction probabilities. The model exhibits the highest confidence in the action recognition with 93.11% in the shooting class and 92.89% in the punching class. This indicates strong certainty in classifying these classes. In contrast, the knife class with 79.51% and the head-hitting class with 71.49% show lower mean confidence, suggesting greater ambiguity in these classes. This is likely due to subtle visual differences or occlusion in video frames. The standard deviations reflect variability across samples, with the kicking class showing the highest variance of 30.98%, indicating that confidence fluctuates more for this action under different scenarios. Overall, these results demonstrate that the model is not only accurate but also provides an interpretable confidence level, which can help prioritize interventions in real-time violence detection.

Table 6
Comparative Performance of WebVidGuard and Baseline Models for Object Detection

Model	Accuracy	Precision	Recall	F1 Score
3D-CNN (Standalone)	87.50%	90.20%	84.10%	87.05%
GMVAE (Standalone)	89.60%	92.00%	87.30%	89.59%
WebVidGuard	90.83%	95%	87.69%	91.20%

Table 6 presented the comparative performance of WebVidGuard and the standalone baseline models for object detection. The results show that WebVidGuard achieved the highest performance across all evaluation metrics, with an accuracy of 90.83% and an F1-score of 91.20% compared to the standalone 3D-CNN and GMVAE models. The hybrid approach demonstrated improved precision of 95%. This indicated a reduced rate of false positives in detecting violent objects such as guns and knives. The integration of 3D-CNN enables effective spatial feature extraction. While GMVAE contributed to improved clustering and classification of object patterns. This combination enhanced the model's ability to distinguish between violent and non-violent objects more accurately than when using either model independently.

Table 7
Comparative Performance of WebVidGuard and Baseline Models for Action Recognition

Model	Accuracy	Precision	Recall	F1 Score
3D-CNN (Standalone)	90.90%	90.10%	89.40%	89.75%
GMVAE (Standalone)	89.60%	88.33%	86.32%	87.31%
WebVidGuard	91.25%	92.50%	90.24%	91.36%

Table 7 presented the comparative performance of WebVidGuard and the standalone baseline models for action recognition. The results indicated that WebVidGuard outperformed both standalone models, achieving an accuracy of 91.25% and an F1-score of 91.36%. The hybrid model also obtained a higher precision of 92.50% and a recall of 90.24%, demonstrating its

effectiveness in correctly identifying violent actions such as punching, kicking, head-hitting, and shooting. Compared to the standalone 3D-CNN model, which focused on spatial-temporal features, and the GMVAE model, which emphasized clustering. The combined architecture leveraged the strengths of both approaches. This integration improved the model's ability to capture temporal dynamics and reduce misclassification, particularly in complex or ambiguous video sequences.

DISCUSSION

The WebVidGuard model performed well in both object detection and action recognition. In object detection, the model demonstrated high accuracy and reliability in identifying violent objects such as guns and knives. For action recognition, the model achieved an F1-score of 91.36%, reflecting strong performance across the four violent action categories: punching, kicking, head-hitting, and shooting. These results indicate that the hybrid approach successfully integrates the 3D-CNN's spatial-temporal feature extraction with the clustering capability of GMVAE, enabling efficient generalization of complex violent actions.

Analysis of the prediction reliability further supports these findings. Confidence level was highest for shooting with 93.11% and punching with 92.89%, while head-hitting with 71.49% and knife with 79.51% exhibited lower confidence and higher variability. The lower performance for head-hitting and knife detection may be attributed to visual similarity with non-violent movements or occasional occlusion, highlighting areas for potential model improvement.

False positives and false negatives in the system have an important connection in detecting violence. False positives may unnecessarily block appropriate content, disrupting the viewing experience, while false negatives could expose children to harmful content. In our evaluation, the hybrid model minimized false negatives relative to standalone baselines, improving safety while maintaining usability. These considerations are critical when designing real-world interventions for content filtering.

The comparative performance of WebVidGuard and the standalone 3D-CNN and GMVAE baseline models for both object detection and action recognition showed that WebVidGuard consistently outperformed the individual models across all evaluation metrics. In object detection, the hybrid model achieved an accuracy of 90.83% and an F1-score of 91.20%, with a high precision of 95%. This indicated a reduced rate of false positives in identifying violent objects such as guns and knives. Similarly, in action recognition, WebVidGuard obtained an accuracy of 91.25% and an F1-score of 91.36%. This demonstrated improved capability in recognizing violent actions such as punching, kicking, head-hitting, and shooting. Compared to standalone models, the hybrid architecture effectively leveraged the spatial-temporal feature extraction of 3D-CNN and the clustering strength of GMVA, resulting in enhanced classification performance and reduced misclassification. These findings confirmed that integrating both models provided a more robust and reliable approach for detecting violent content in videos across multiple dimensions.

Overall, WebVidGuard model using 3D-CNN + GMVAE demonstrates the most consistent performance across all dimensions, including detection accuracy, temporal context, speed impact, and prediction confidence. Its combination of high precision, reliable temporal modeling, and superior adaptability suggests that the hybrid 3D-CNN + GMVAE architecture provides a comprehensive and practical solution for real-world violent video detection.

Conclusion. This study developed and evaluated WebVidGuard, a real-time web-based video violence detection system integrating 3D-Convolutional Neural Network (3D-CNN) and Gaussian Mixture Variational Autoencoders (GMVAE). Across five evaluation dimensions, the system demonstrated high accuracy, precision, recall, and F1 score performance. The

system outperformed baseline models that used standalone 3D-CNN or GMVAE components for video violence detection task

Among the four violent action categories, shooting achieved the highest detection performance in the conducted experiments. Head-hitting was the most challenging due to visual similarity with non-violent movements and occasional occlusion. Comparative analysis with prior studies indicated that the hybrid 3D-CNN + GMVAE architecture improved detection performance. The model effectively captures spatial-temporal patterns and reduces false positives and false negatives in video analysis.

These findings confirm that integrating spatial-temporal feature extraction with unsupervised clustering enhances the accuracy and robustness of video violence detection. The system can be implemented in web environments in real-time, offering practical applications in parental content filtering and online safety for children.

Limitations of the Study. Even with the promising results, the study has several limitations. First, the efficiency of the model depends largely on the quality and diversity of the training dataset. Variations in lighting, occlusion, and camera angles may affect detection performance. Second, the model showed lower confidence and reliability in certain action classes, such as head-hitting, indicating challenges in distinguishing subtle or ambiguous movements. Finally, real-time deployment performance may vary depending on hardware capabilities and network conditions.

Future Direction. Future researchers should look into using new technologies, like WebAssembly, HTML5 video APIs, or other upcoming methods, and integration with modern streaming platforms to enhance real-time content filtering to make the app even better at finding violent content. Future research could try combining different technologies, for example, they could use YOLO for spotting objects, RNN for understanding how things change over time, and WebVidGuard itself. They could also add audio analysis (like listening to screams or explosions) alongside video analysis, which could enhance detection accuracy, especially in cases where violent actions are partially obscured or occur outside the visual frame.

Author contributions. Rengel V. Corpuz: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Resources; Software; Visualization; Writing – original draft; Writing – review & editing | Aleta C. Fabregas: Supervision; Validation.

Conflict of interest. The authors declare no conflict of interest.

Funding source. This research received no external funding.

Artificial intelligence use. AI-assisted language editing was performed using Grammarly; authors reviewed and approved all content.

Ethics approval statement. Ethical approval was obtained from the Institutional Review Board, with reference code 2025-136.

Data availability statement. The datasets generated and/or analyzed during the current study are available in the Kaggle repository, https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset/data.

Acknowledgement. First and foremost, all glory and honor are given to Almighty God, whose guidance, wisdom, and strength have been the foundation on this journey. Without His grace, this research would not have been possible.

Sincere gratitude is extended to Ms. Marife Corpuz for her constant love, support, and understanding. Her encouragement served as a source of strength throughout the challenges of this research. Appreciation is also extended to RM Rich Corpuz for his boundless joy and inspiration that served as motivation in completing this study.

Heartfelt appreciation is extended to Dr. Aleta C. Fabregas for her invaluable guidance, patience, and continuous support throughout the conduct of this study. Her expertise, encouragement, and constructive feedback greatly helped shape the direction and quality of this research. Gratitude is also expressed to the panel of experts, headed by Dr. Benilda Eleonor V. Comendador. Together with the esteemed panel members, Dr. Remedios G. Ado, Dr. Rosicar E. Escober, and Prof. Leoven R. Basista. Their insightful comments, suggestions, and professional critiques significantly improved this study.

Gratitude is likewise extended to the faculty members and the institution for providing the knowledge, resources, and academic environment that made this research possible. Finally, sincere thanks are given to family and friends for their unwavering moral support, encouragement, and understanding throughout the entire research journey. Above all, gratitude is once again offered to Almighty God for the wisdom, strength, and guidance bestowed in the completion of this work.

REFERENCES

Aremu, T., Zhiyuan, L., Alameeri, R., Khan, M., & Saddik, A. El. (2023). *SSIVD-Net: A Novel Salient Super Image Classification & Detection Technique for Weaponized Violence*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2207.12850

Cheng, M., Cai, K., & Li, M. (2020). *RWF-2000: An Open Large Scale Video Database for Violence Detection*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1911.05913

Escobanez, J. C. S., & Comendador, B. E. (2022). Student Physical Violence Detection using Convolutional Neural Networks. In *ICICM 2022: The 12th International Conference on Information Communication and Management*, 34–38. ACM. https://doi.org/10.1145/3551690.3551696

Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M. D., & Xiao, F. (2020). Video anomaly detection and localization via Gaussian Mixture Fully Convolutional Variational Autoencoder. *Computer Vision and Image Understanding, 195*. https://doi.org/10.1016/j.cviu.2020.102920

Fitzpatrick, C. (2018). *Watching violence on screens makes children more emotionally distressed*. The Conversation. https://doi.org/10.64628/AAJ.mrx9qsxh9

Gao, H. (2023). A Yolo-based Violence Detection Method in IoT Surveillance Systems. *International Journal of Advanced Computer Science and Applications, 14*(8). https://doi.org/10.14569/IJACSA.2023.0140817

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org/

Guedes, A. R. M., & Chavez, G. C. (2020). Real-Time Violence Detection in Videos Using Dynamic Images. In *2020 XLVI Latin American Computing Conference (CLEI)*, 503–511. IEEE. https://doi.org/10.1109/CLEI52000.2020.00065

Hailesellasie, M. T., & Hasan, S. R. (2019). MulNet: A Flexible CNN Processor with Higher Resource Utilization Efficiency for Constrained Devices. *IEEE Access, 7*, 47509–47524. https://doi.org/10.1109/ACCESS.2019.2907865

Hazimah binti Wan Ismail, W., Ramadhani Mohd Husny, H., Syahman bin Mamat, A., & Ya Abdullah, N. (2020). Parental Control System for Children on Wireless Network. *Journal of Computing Technologies and Creative Content, 5*(1).

Jain, A., & Vishwakarma, D. K. (2020). State-of-the-arts Violence Detection using ConvNets. *2020 International Conference on Communication and Signal Processing (ICCSP)* (pp. 0813–0817). IEEE. https://doi.org/10.1109/ICCSP48568.2020.9182433

Juneja, A., Juneja, S., Soneja, A., & Jain, S. (2021). Real time object detection using CNN based single shot detector model. *Journal of Information Technology Management, 13*(1), 62–80. https://doi.org/10.22059/jitm.2021.80025

Luo, H., Xie, W., Wang, X., & Zeng, W. (2019). *Detect or Track: Towards Cost-Effective Video Object Detection/Tracking*. arXiv. https://doi.org/10.48550/arXiv.1811.05340

Khan, S. U., Haq, I. U., Rho, S., Baik, S. W., & Lee, M. Y. (2019). Cover the violence: A novel deep-learning-based approach towards violence-detection in movies. *Applied Sciences (Switzerland), 9*(22). https://doi.org/10.3390/APP9224963

Mani, V. R. S., Saravanaselvan, A., & Arumugam, N. (2022). Performance comparison of CNN, QNN and BNN deep neural networks for real-time object detection using ZYNQ FPGA node. *Microelectronics Journal, 119*, 105319. https://doi.org/10.1016/j.mejo.2021.105319

Muhammad, K., Mustaqeem, Ullah, A., Imran, A. S., Sajjad, M., Kiran, M. S., Sannino, G., & de Albuquerque, V. H. C. (2021). Human action recognition using attention-based LSTM network with dilated CNN features. *Future Generation Computer Systems, 125*, 820–830. https://doi.org/10.1016/j.future.2021.06.045

Nwokonkwo, O., Samuel, N., Eze, U., & John-Otumu, A. (2024). Machine Learning Framework for Real-Time Pipeline Anomaly Detection and Maintenance Needs Forecast Using Random Forest and Prophet Model. *Automation, Control and Intelligent Systems, 12*(2), 22–34. https://doi.org/10.11648/j.acis.20241202.11

Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., Stringhini, G., & Sirivianos, M. (2020). Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children (Vol. 2020). *Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)*. https://doi.org/10.48550/arXiv.1901.07046

Park, J. H., Mahmoud, M., & Kang, H. S. (2024). Conv3D-Based Video Violence Detection Network Using Optical Flow and RGB Data. *Sensors, 24*(2). https://doi.org/10.3390/s24020317

Peixoto, B., Lavi, B., Pereira Martin, J. P., Avila, S., Dias, Z., & Rocha, A. (2019). Toward Subjective Violence Detection in Videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8276–8280. IEEE. https://doi.org/10.1109/ICASSP.2019.8682833

Sabokrou, M., Fayyaz, M., Fathy, M., & Klette, R. (2017). Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Image Processing, 26*(4), 1992–2004. https://doi.org/10.1109/TIP.2017.2670780

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Zahra., & Klette, R. (2018). Deep-anomaly: Fully convolutional neural

network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding, 172*, 88–97. https://doi.org/10.1016/j.cviu.2018.02.006

Sangwan, D., & Jain, D. K. (2019). An evaluation of deep learning-based object detection strategies for threat object detection in baggage security imagery. *Pattern Recognition Letters, 120*, 112–119. https://doi.org/10.1016/j.patrec.2019.01.014

Shutsko, A. (2020). User-generated short video content in social media: A case study of TikTok. In: Meiselwitz, G. (eds) *Social computing and social media: Participation, user experience, consumer experience, and applications of social computing*. HCII 2020. (Vol. 12195, pp. 108–125). Springer. https://doi.org/10.1007/978-3-030-49576-3_8

Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., 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. IEEE. https://doi.org/10.1109/ICICIS46948.2019.9014714

Song, W., Zhang, D., Zhao, X., Yu, J., Zheng, R., & Wang, A. (2019). A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks. *IEEE Access, 7*, 39172–39179. https://doi.org/10.1109/ACCESS.2019.2906275

Tamam, Moh. B., Hozairi, H., Walid, M., & Bernardo, J. F. A. (2023). Classification of Sign Language in Real Time Using Convolutional Neural Network. *Applied Information System and Management (AISM), 6*(1), 39–46. https://doi.org/10.15408/aism.v6i1.29820

Ullah, F. U. M., Ullah, A., Muhammad, K., Haq, I. U., & Baik, S. W. (2019). Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors, 19*(11). https://doi.org/10.3390/s19112472

Vinney, C. (2023). *How violent media can impact your mental health*. Verywell mind. https://www.verywellmind.com/what-is-the-impact-of-violent-media-on-mental-health-5270512

Yao, G., Lei, T., & Zhong, J. (2019). (A). A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognition Letters, 118*, 14–22. https://doi.org/10.1016/j.patrec.2018.05.018

Yao, G., Lei, T., Zhong, J., & Jiang, P. (2019). (B). Learning multi-temporal-scale deep information for action recognition. *Applied Intelligence, 49*(6), 2017–2029. https://doi.org/10.1007/s10489-018-1347-3

Zhang, T. (2020). A Brief Study on Short Video Platform and Education. In *2nd International Conference on Literature, Art and Human Development (ICLAHD 2020)* (pp. 264–267). Atlantis Press. https://doi.org/10.2991/assehr.k.201215.494