Deep Learning based Automatic Violence Detection system for Efficient Monitoring of Violence Incidents

Sinchana R Naik[1],Spandhana[2],Sushmitha E[3],Veena G T[4],Mr. Giridhar Gowda[5]
[1]Alva's Institute of Engineering and College,VTU,Moodbidri,India, sinchananaik62@gmail.com
[2]Alva's Institute of Engineering and College,VTU, Moodbidri,India, spandhanan21@gmail.com
[3]Alva's Institute of Engineering andCollege,VTU,Moodbidri,India,4al20cs156@aiet.org.in
[4]Alva's Institute of Engineering and College,VTU,Moodbidri,India,veenagowda9893@gmail.com
[5]Alva's Institute of Engineering and College,VTU, Moodbidre, India,giridhargowda@aiet.org.in

I. INTRODUCTION

Throughout the world, CCTV camera widely used and Possess Ability to provide extensive and reliable information for a various type of security applications [1]. But the capacity to make judgements quickly which is crucial in employing video monitoring to detect crime and violence is compromised by the requirement to examine hours of video material [2]. To decrease the burden to authorities who are watching hours of footage to identify brief incidents, a number of research regarding the automatic recognition of violent sequences in films have been published. Spatial-temporal features that is, characteristics that convey both the spatial information in a single frame and the motion information containedd in a sequence of frames can be extracted from films. using deep learning approaches, as proven by studies[3]. Among the most successful deep learning-based methods for detecting violence is the combination of 3D Convolutional Neural Networks (3D CNN) and Support Vector Machines (SVM), which we demonstrated in a previous study to identify both interpersonal and group violence in films.[1] We indicated that there are now false positives, which interpret fast or friendly actions like high fives, embraces, tiny hits, and claps as aggressive. To explore this further, this work compares three distinct deep learning models on the violence dataset [3], a unique dataset that we constructed to contain video clips that may result in false positives as non-violent examples. In particular, this work advances the field of automated violence identification in the subsequent ways. The possibility of extracting features from pretrained models such as VGG-16,

VGG-19, and ResNet-50 is investigated in this work, which delves deeply into transfer learning. Furthermore, in one of the study's tests, we sent 30 frames of a video at a time into the LSTM network to take use of its memory and forgetting capabilities.

II. LITERATURE REVIEW

[1]. Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network by Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq and Sung Wook Baik. This paper proposes a triple-staged deep learning framework for identifying violence in smart city surveillance systems. It employs lightweight CNNs for person detection, 3D CNNs for extracting spatiotemporal features, and utilizes an optimization toolkit for efficient model execution. Upon detecting violent activity, alerts are sent to nearby authorities for prompt intervention, outperforming existing methods on benchmark datasets.

[2]. Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset by PAOLO SERNANI, NICOLA FALCIONELLI, SELENE TOMASSINI, PAOLO CONTARDO, AND ALDO FRANCO DRAGONI This study introduces three deep learning models for violence detection in videos, tested on the AIRTLab dataset designed to challenge algorithms against false positives. Results demonstrate the efficiency of transfer learning-based networks over those trained from scratch, with most errors attributed to misidentifying non-violent clips, validating the dataset's design. Additionally, comparison with existing literature and 2D CNN-based models underscores the superior performance of 3D models in processing spatiotemporal features, with experiment code and the dataset available for public access.

[3]. Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications by VIKTOR

DÉNES HUSZÁR IMRE NÉGYESI, VAMSI KIRAN ADHIKARLA, AND CSABA KRASZNAY. the use of 3D convolutional neural networks and pre-trained action recognition models for automatic violence detection in surveillance footage. By extending and evaluating on diverse datasets, the performance of the suggested approach is better than state-of-the-art approaches, attaining greater precision with less parameters and robustness against common compression artifacts.

[4]. Violence Detection by Pretrained Modules with Different Deep Learning Approaches by Shakil Ahmed Sumon, Raihan Goni. This paper investigates several approaches for identifying salient features in pretrained models such as VGG16, VGG19, and others, violence detection from video ResNet50. By feeding features extracted from ResNet50 into an LSTM network, combined with attention mechanisms, the study achieves superior accuracy of 97.06%, outperforming other models explored in the experiment.

III. RELATED WORKS

Deep learning models have demonstrated their potential on the aforementioned datasets pretraining to violence detection approaches in recent times, demonstrating exceptional performance regarding classification accuracy. Of these methods, 3D CNNs and ConvLSTMs have to be successful in acquiring the spatiotemporal knowledge seen in videos. Regarding the prediction of violence, several academics have used an alternative methodology. Four distinct feature types, including audio features, attribute features, trajectory-based motion features, and spatial-temporal interest points (STIP), were utilized into the construction of a novel model. The interest spots have been found using the STIP method in both temporal and geographical dimensions. Among the features is the audio that is extracted from every video[2]. The support vector machine is utilized for classification (SVM). With this model, an accuracy of 68.2% has been achieved.

The 246 videos (123 violent and 123 non-violent) that were acquired from YouTube with an average length of 3.6 seconds and a 320 x 240 of resolution make up the Crowd Violence Dataset [1]. The Hockey Fight and Crowd Violence datasets contain footage shot in extremely specific locations (such football stadiums and hockey arenas) all of these datasets contain poor resolution movies.

The Crowd Violence Dataset, the Movie Fight Dataset, and the Hockey Fight Dataset were the datasets that are often used for comparing violence detection systems. There are 1000 clips in the Hockey Fight Dataset, split evenly between fight and no fight. Although the 320 x 240-pixel version as suggested in is frequently utilized, each clip comprises between 41 and 50 frames (as also mentioned in), initially at a size of 720 × 576 pixels.

The model made advantage of spatiotemporal characteristics and local optical flow method. The optical flow approach and the Harris 3D spatial-temporal interest point detector were merged in the model's creation. However, they had a highly inconsistent outcome their best accuracy was 69.43%. These days, a lot of people own security cameras. In light of this, scientists attempted to create an enhanced violence detector by taking two distinct actions. First, a feature extraction technique was created with an emphasis on variations in motion magnitude. Next, they made an effort to include characteristics and multiclassification in the nation. They got fantastic results when they ran their models on two available datasets.

IV. IMPLEMENTATION

The suggested system is put into practice by doing the following actions.

Pre-processing

Data Pre-processing is a methodology that is used to convert the raw data into a clean data set. In other words, whenever the information is gathered from different sources it is collected in raw format which is not feasible for the analysis. Dataset is pre-processed to resize the dataset images. By using user data, produced our own dataset where images are resized to 150*150 before processing them further.

Train Test Split

The process of organizing data into groups and classes on the basis of certain characteristics is known as the classification of data. Classification helps in making comparisons among the categories of observations. It can be either according to numerical characteristics or according to attributes. So here we need to visualize the prepared data to find whether the training data contains the correct label, which is known as a target or target attribute. In this project the dataset is split into 90 samples for training and 10 for testing.

Training

The methodology of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model means the model artifact that is created by the training process. The training data must contain the correct answer, which is referred to as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

Picking the Model

Pickling is a process in which model is stored in a file for future use. Before pickling model must be trained well and optimized to maximum extent possible because after picking model, data will not be trained. By pickling there is no need

of training the model each time the user makes a classification.

3D CNN

A type of deep learning architecture called three-dimensional CNNs is employed for video analysis jobs that call for both spatial and temporal data. An effective neural network architecture for processing three-dimensional data, like video or volumetric images, it has the ability to record both spatial and temporal information is called a 3D CNN, or 3D Convolutional Neural Network. With dimensions denoting width, height, and time (frames), video data is inherently three-dimensional. In a video, every frame is seen as a "slice" of the 3D volume. From unprocessed video data, 3D CNNs may extract hierarchical features.

Deeper layers record more intricate spatiotemporal patterns, while lower layers may catch basic elements like edges and textures. Video sequences can be inputted into 3D-CNN, that is a 2D-CNN extension. In order to take out spatiotemporal data from a video sequences, the 3D CNN architecture typically consists of many convolutional layers and pooling layers[6].

The final prediction is then produced by passing the result of the convolutional layers via fully linked layers and activation functions. Action recognition, gesture recognition, and video-based violence detection are only a handful of the video analysis tasks in which 3D CNNs have shown effective. 3D CNNs are able to accomplish cutting edge results on these tasks by utilising both spatial and temporal information, particularly when working with dynamic and intricate films.

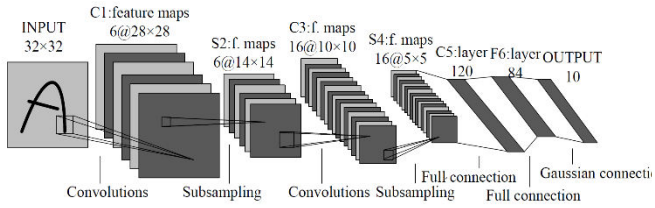

Figure 1. Convolutional Neural Network

video analysis tasks in which 3D CNNs have shown effective. 3D CNNs are able to accomplish cutting edge results on these tasks by utilising both spatial and temporal information, particularly when working with dynamic and intricate films.

A Multiple instance Learning (MIL)-based method that trains a 3D Convolutional feature from different video segments only the prototype with video-level labels is based on a fully-connected neural network structure. Next, for each positive bag (which contains aberrant movies) and negative bag (which contains normal videos), the network's performance was analysed between the highest and lowest-scoring instances using a remarkable ranking loss method[6]. Two-stream is another application of 3D-CNN. In VD tasks, CNN is a popular deep learning architecture. Due of its capacity to record both spatial and temporal information, this technique rose to fame. Using two distinct streams a spatial stream that gathers motion information from the frames and a temporal stream that pulls static appearance information this method includes processing video frames. In order to extract appearance features, the spatial stream feeds a CNN architecture with the raw values of RGB pixel from frame. In contrast, the temporal stream calculates optical flow from the frames and feeds it into a different CNN to extract motion properties. Ultimately, a final forecast is produced by merging the output features from both streams.

The videos have been scaled to 28 by 28 pixels for our CNN architecture because they are of varying resolutions. To Accelerate the training, we have reduced the frame's quality. From the videos, a total of 30 frames are taken out per second.[3] The architecture of a convolutional neural network (CNN) is layered. It usually consists of an input layer, a few pooling and convolution layers together, and an output layer. CNN serves as an extractor of features. Following that, the features are sent via a few fully connected layers, which work together to create a classifier. the CNN design we have presented consists of three densely linked layers, a sigmoid layer, and two convolution layers. Each of the dense layers contains ten nodes, whereas the convolution layers have 32 filters. The filters utilized in the convolution layers have a 3X3 size. After the convolution and fully linked layers, there is a batch normalization layer. Furthermore, we have retrained a pretrained model that we first trained using a dataset of violent films using the movies that we have gathered. The CNN model's design and that of the pretraining model are comparable. The movie dataset provided the model with attributes that it learned, and these features helped it distinguish between aggressive and passive movies in the violence dataset. The 224 x 224-pixel pictures that are fed to these pretrained models must meet certain requirements. We then sent them to the models. Following a sigmoid output layer, these characteristics are sent into three fully linked layers for classification.

To accomplish an input-to-output mapping, it functions fundamentally as a multi-layer perceptron that mimics local perception. Several convolutions and pooling are used to extract the properties of the data at various scales. The way local connections and shared weights are handled in the CNN network is distinctive. Reducing the number of weights not only makes the network easier to improve, but it also lowers the chance of overfitting. The three mutually supporting layers that make up a CNN are the convolutional layer, pooling layer, fully connected layer, and Softmax layer. We obtain local features throughout the convolution procedure.

This process requires obtaining more complex feature correlation values from low-level convolutional layers

through multi-level cascading since among the convolution layers is made up of multiple convolution units. This is necessary to extract more features about the input parameters.

A. System Architecture

User will communicate with the system using web application. Every uploaded test video will be processed by the deep learning model and predict its class. Deep learning module has four main functionalities such as create model, train model, saving the model weight and classification. Dataset needed for the proposed system is obtained from the Internet and augmented to create more copies using different augmentation methods such as scaling, zooming and rotation. A CNN model is built to train the system by extracting image features. DL model predicts the type of violence category based on test video.

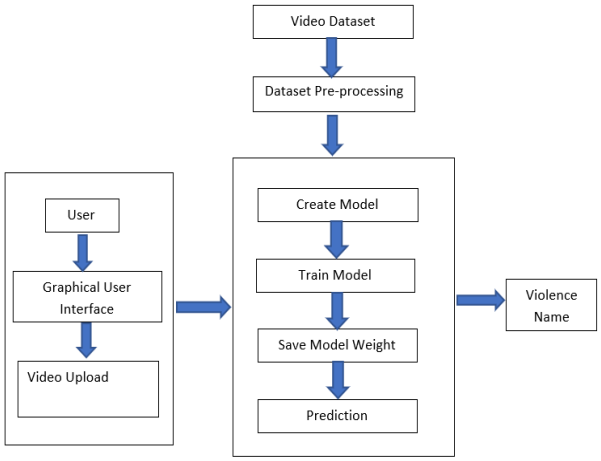

Figure 2. System Architecture

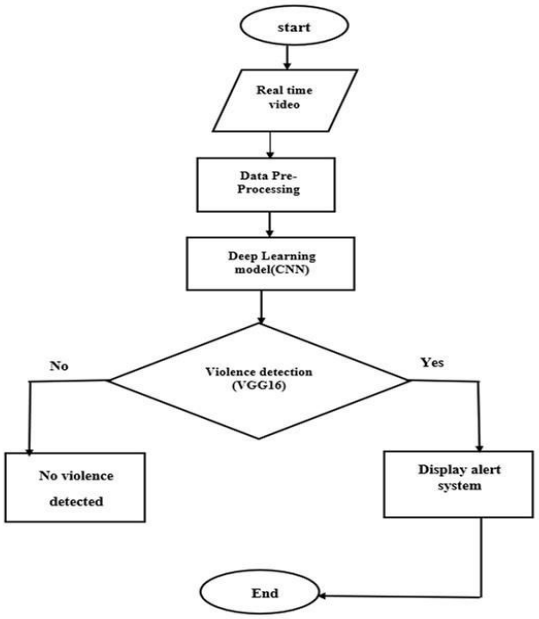

Figure 3. Flowchart of violence Detection System

In our system the real time videos are collected from surveillance camera and it is sent to data pre-processing where we get a relevant frames. The frames are passed to 3D convolutional neural networks (3D CNNs) to capture temporal dependencies in the video data. If there is a violence then the alert message will be displayed if not it takes the other frames.

B. Software Used

Python

Python is a high-level, interpreted, general-purpose programming language. Codability is prioritized in its design philosophy with the use of significant indentation.

The violence detection model can be trained with data collected and preprocessed using Python. This could entail obtaining picture or video data from multiple sources, labeling the data (for example, dividing up the frames into violent and non-violent categories), and preparing the data.

Deep learning architectures for violence detection can be defined and implemented by developers using Python. To evaluate video sequences or image frames, neural network designs such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or more sophisticated architectures like convolutional long short-term memory networks (ConvLSTMs) are designed.

Django – Python Web Framework

Django is a Python-based free and open-source web framework that follows the model–template–views (MTV) architectural pattern. It is maintained by the Django Software Foundation (DSF), an independent organization established in the US as a non-profit.

The violence detection system can be made to have a web interface with Django, so that users can interact with it via a browser. Features like checking analysis findings, adjusting system settings, and uploading movies for analysis might all be included in this interface.

Keras

Keras is an open-source software library that provides a Python interface for artificial neural networks.

For creating deep learning models, Keras offers an easy-to-use API. With Keras's sequential or functional API, developers may quickly construct a neural network architecture appropriate for violence detection. Determining the neural network's layers entails specifying its convolutional layers for image processing and its recurrent layers for sequence analysis, if relevant.

Using labeled data, developers can use Keras to train deep learning models for violence detection. The fit() function in Keras can be used to feed the model with training data, which

is usually pictures or video frames that have been classified as violent or non-violent. Training parameters that developers can set include the number of epochs, batch size, and optimization.

HTML

Hypertext Markup Language (HTML) is the standard markup language for creating web pages and web applications. With Cascading Style Sheets (CSS) and JavaScript, it forms a triad of cornerstone technologies for the World Wide Web.

Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the document.

The user interface (UI) of the violence detection system is made with HTML. This includes creating websites where people may view the outcomes of the violence detection model, upload videos or photographs for analysis, and engage with the system.

Users can upload photos or video files for the deep learning model to examine through the user interface (UI) using HTML form components such file input fields. JavaScript can be used to initiate file upload events when a user selects a file, sending the data to the server for processing.

JavaScript

Web application UIs are frequently created using JavaScript. JavaScript may be used to develop an intuitive interface for users to upload or stream video feeds for analysis in the context of violence detection. Developers may execute pre-trained deep learning models directly in the browser with frameworks like TensorFlow.js. This eliminates the need to transport data to a distant server for processing, enabling real-time video processing and analysis. JavaScript may be used to visualize the findings of violence detection. For example, it can be used to highlight specific regions in a video where violence is identified or to provide statistics on the frequency and severity of violent behaviour.

Back End: MY-SQL

A violence detection system that uses a deep learning model can employ MySQL, a relational database management system, to store several kinds of data connected to the system's operation, such as user information, submitted videos, analysis results, and system logs.

MySQL is used to store user data, such as email addresses, hashed passwords for security, and any other pertinent user attributes. By doing this, the system is able to manage user access permissions to its features and functionalities and authenticate users.

C. Operating System: Microsoft Windows XP/Windows 7

Windows XP/7analyses the performance impact of visual effects and uses this to determine whether to enable them, so as to prevent the new functionality from consuming excessive additional processing overhead. Users can further customize these settings. Some newer video cards. However, if the video card in not capable of hardware alpha blending, performance can be substantially degraded, and Microsoft recommends the feature should be turned off manually. Windows XP/7 added the ability for windows to use “Visual Styles” to change the appearance of the user interface.

V. Expected Outcome

To precisely locate violent scenes in video feeds, a deep learning-based automatic violence detection system is being developed. To achieve reliability, the model should be capable to recognize violent acts reliably while limiting false positives. When violence is detected, the system need to be capable to quickly notify the appropriate authorities or individuals so that they can take appropriate action or respond. To be able to quickly detect and address violent occurrences, it should be able to process and analyse media streams in real-time.

To be able to ensure that violent actions appear realistic, the system should be resilient to changes in background colors, lighting, camera angles, and other environmental elements.

To manage massive amounts of media data possibly from several sources at once it should be scalable. The system should be flexible enough to accommodate diverse forms of violence, such as physical altercations, aggressive conduct, or the utilisation of weapons, and intelligent enough to identify these occurrences in a range of situations. Making certain system respects people's right to privacy by refraining from interfering with their private lives or spaces unless absolutely required for safety or security.

Overall, improving public safety and security in a diversity of settings, such as public areas, transit hubs, workplaces, and schools, would be made possible by the effective deployment of an automatic violence detection system utilizing deep learning.

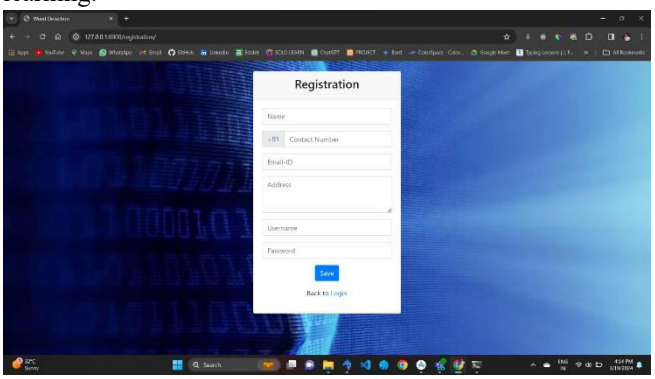

Figure 4 Registration Page

This is the registration page for the user. If the user is new to the page that they must signup by giving their credentials.

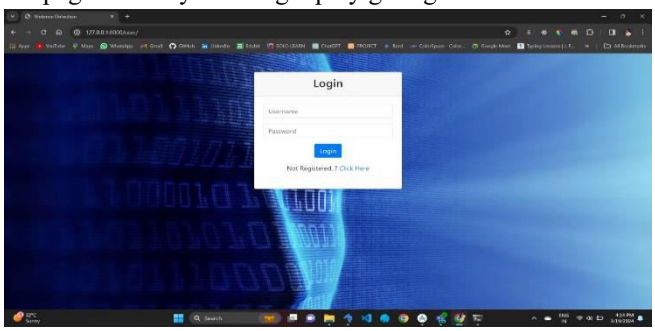

Figure 5 Login Page

This is the login page where the user who have already registered can give their username and password for logging in.

Figure 6. Violence detection

Figure 7. Violent and Non violent classification

VI. CONCLUSION

In summary, the use of a deep learning-based automated violence detection system has enormous capacity to get better public safety and security in a diversity of contexts. Such a system can efficiently analyse live video streams to acknowledge and categorize violent behaviour with an excellent level of precision by utilizing deep neural networks and sophisticated algorithms.

By integrating deep learning models, the system can recognize complex patterns and characteristics that are representative of violent conduct, which helps it differentiate between aggressive and normal behaviours in a variety of contexts. The model can capture temporal and spatial relationships within video data, providing robust and reliable detection capabilities, by utilizing the power of convolutional neural networks (CNNs), recurrent neural networks (RNNs), or hybrid architectures.

Furthermore, by enabling prompt responses and more effective threat mitigation, the implementation of such automated violence detection systems might greatly improve the capacities of law enforcement and security professionals. These systems not only offer real-time monitoring but also, through data analytics, can offer insightful information that can help design proactive plans for risk mitigation and crime prevention.

The utilization of such systems may raise ethical and privacy issues, though, which should be recognized and addressed. These concerns may include those pertaining to algorithmic transparency, prejudice, and surveillance. To guarantee the proper and ethical utilizing new technologies, it is essential to strike a balance between security imperatives and individual rights.

In conclusion, deep learning-based automated violence detection systems are an effective instrument for improving public safety, but their creation, application, and use need to be directed by legal and ethical frameworks to protect privacy, equity, and responsibility. These systems have the potential to significantly contribute to making all surroundings safer and more secure with sustained study, innovation, and cooperation.

REFERENCES

[1]. PAOLO SERNANI, NICOLA FALCIONELLI1, "***Deep Learning for Automatic Violence Detection:Tests on the AIRTLab Dataset"***

[2] Shakil Ahmed Sumon, Raihan Goni, ***"Violence Detection by Pretrained Modules with Different Deep Learning Approaches"***

[3] Prof. R.U. Shekokar, Lokesh Padme, "***VIOLENCE DETECTION USING DEEP LEARNING***"

[4] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq and Sung Wook Baik "***Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network"***

[5]. VIKTOR DÉNES HUSZÁR IMRE NÉGYESI, VAMSI KIRAN ADHIKARLA, AND CSABA KRASZNAY ***"Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications"***

[6] SOHEIL VOSTA, (Student Member, IEEE), AND KIN-CHOONG YOW ***"KianNet: A Violence Detection Model Using an Attention-Based CNN-LSTM Structure"***

[7] MAHMOUDABDELKADER BASHERY ABBASS AND HYUN-SOO KANG ***"Violence Detection Enhancement by Involving Convolutional Block Attention Modules Into Various Deep Learning Architectures: Comprehensive Case Study for UBI-Fights Dataset***"

[8]. A Hanson, K. PNVR, S. Krishnagopal, and L. Davis, ***''Bidirectional convolutional LSTM for the detection of violence in videos,''*** presented at the Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018.

[9]. Bilinski and F. Bremond, ***''Human violence recognition and detection in surveillance videos,''*** in Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveillance (AVSS), Aug. 2016, pp. 30–36.

[10]. PenzeyMoog and D. C. Slakoff, ***''As technology evolves, so does domestic violence: Modern-Day tech abuse and possible solutions,''*** in The Emerald International Handbook of Technology-Facilitated Violence and Abuse. Bingley, U.K.: Emerald Publishing Limited, 2021.

[11]. W Shin, S.-J. Bu, and S.-B. Cho, ***''3D-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance,''*** Int. J. Neural Syst., vol. 30, no. 6, Jun. 2020, Art. no. 2050034.

[12]. A Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, ***''Large-scale video classification with convolutional neural networks,''*** in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1725–1732.

[13]. D Harkin and R. Merkel, ***''Technology-based responses to technology facilitated domestic and family violence: An overview of the limits and possibilities of tech-based 'solution',''*** Violence Against Women, vol. 29, nos. 3–4, pp. 648–670, Mar. 2023.H