Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Enhanced Violence Detection Using Convolution Neural Networks Approach

Gokila Deepa G [1], G Dhivyasri [2*], Manaswini R [3], Remya R [4], M Manikandan [5], Padmavathi M [6]
[1] *Dept. of AI&DS, PPG Institute of Technology, Tamil Nadu, India.*
[2] *Dept. of CSE (Data Science), Sai Vidya Institute of Technology, Karnataka, India.*
[3,5] *Dept. of Electronics and Communication Engineering, Presidency University, Karnataka, India*
[4] *Dept. of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Tamil Nadu, India*
[6] *Dept. of Electronics and Communication Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*
* ***Corresponding Author:*** *dhivyaasrigopal@gmail.com*

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Violence detecting through cameras is to find specific object identification and human action is prominent in office sectors and public places to enhance safety. CCTV images contain different objects, different backgrounds, real-time incident in office sectors and public places, in the existing methods residual network and K-nearest neighbors' method were used to classify and detect violence. The disadvantages of these methods are finding objects and human behavior. The accuracy level is also the most complex to achieve. Therefore, in the proposed method, violence detection and object identification using the Inception-v3 architecture and networking techniques are used to improve the accuracy. The Inception-v3 architecture is a convolutional neural network (CNN) that uses a variety of techniques to improve image classification performance. The networking techniques used in research consist of five stages: video input, image augmentation, image classification, email notification, and sound alarm. The video clips from various open-access databases are utilized and converted with different data sets of testing images. Image augmentation improves the blur and saturation of each testing image and makes the testing images better. Yolo v3 algorithm separates the different objects, like the hand, table, and specific parts, for detecting objects. The Inception-v3 architecture classifies each human action for the identification of human behavior. The training images classify each frame of the image dataset, which employs deep learning methods with computer vision technologies. The experiments on these database results with 99.15% accuracy, along with email notification alerts and a sound alarm.

Keywords: Violence detection, Object identification, Inception-v3, Image augmentation, Yolo v3, KNN, Video classification

1. INTRODUCTION

Validation data introduces fresh data that the model still needs to be evaluated during training. Data scientists may assess model predictions based on the new data by using validation data, which offers the first test against new data. Digital image-processing technologies have significantly improved this method of recording captured video streaming and framing them via various networking and mobile application development platforms.

The input data that corresponds to an expected outcome is fed into the Algorithm by the data analyst. The model continuously assesses the data to get a more thorough understanding of the data's behavior, after which it modifies itself to fulfill its intended function. So, because security has been given top importance for people's safety, public places such as shopping malls, roadways, and highways are increasingly being installed with CCTV cameras.

In most cases, violence detection is used to discover exciting facts on targets in clips from videos [1]. The ideal quality of the target object's interest spots is achieved by lighting fluctuations and shifting the camera perspective.

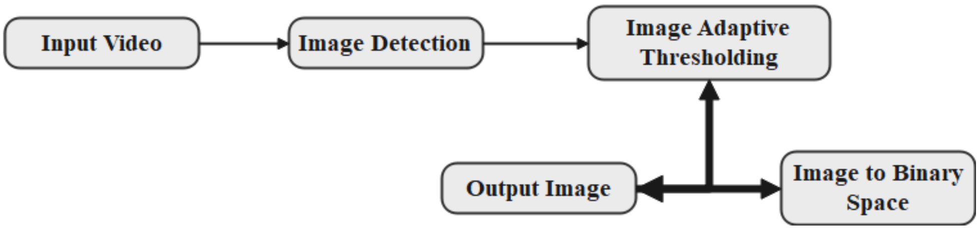


Fig 1. The Analysis of Images Based on Recognition of Movement and Position

Methods for detecting object tracking have two definitions: (i) identifying a moving background item or several objects over time with a camera, and (ii) predicting the object's trajectory in the frame as it travels across a video timeline. While image preprocessing processes are done to both the training and test sets, image augmentation changes are only applied to the training data. During the super-resolution stage, many low-resolution photos are combined to create a high-resolution image, and the workflow is shown in Fig. 1.

They determine whether the recorded activities are aberrant or suspicious and demand the workforce's constant attention. Consequently, this shortcoming is driving a need for highly accurate automation of this operation. Furthermore, it is necessary to show which frames and sections of the video include the unexpected behavior to determine quickly whether or not it is unusual or suspicious. This is a crucial distinction between the two types of image preparation. Therefore, a change that could be best suited as an initial processing phase in certain circumstances might be an augmentation elsewhere.

Subsequently, a dependent object tracking module is started, assisted by the detected object data, identifies each detected object by its unique ID number, and forecasts its future position. However, the quantity of tracking BBox matches the number of detected items if the past tracked BBox is 0. The growing need for increased safety and security in a variety of contexts necessitates the development of a violence detection project applying human action recognition.

The work has the potential to reduce violence and speed up emergency response times by creating precise and trustworthy techniques for identifying aggressive conduct in real time. For instance, if u = 0 in time Tc, then = n'. Stated differently, video footage is typically analyzed to find patterns of human behavior linked to violent acts. Conventional computer vision techniques, such as feature extraction and machine learning algorithms, are used in this manner.

Object identification is essential in a real-time context because it enables various vital applications, such as security and surveillance, which use visual data to identify individuals and provide additional assurance. Mid-level attribute: the feature uses a portion of the pixel to address important structure data, such as an edge, corner, interest point, or constant location, while the entirety of the pixel imparts standard qualities.

1.1. Objective of this research

1. To detect and extract objects, e.g., tables, lights, and fans, from a preprocessed image by using data augmentation

2. To improve edge detection techniques, identify the angles and boundaries in the image and reduce the complexity of each layer of data.

3. To detect human unknown activities using Inception-v3 and categorize the extraction accuracy of each dataset.

4. To detect and classify each data set of images and notifications using deep learning and networking techniques.

2. EXISTING WORK

Bhatti et al. (2023) [2] Re-identification of one specific vehicle dataset was collected from the Campus's Surveillance Cameras (CCTV)—data preprocessing: boundary detection of shots. The value of processing each frame for vehicle reactivation can be reduced by applying a shot boundary detection technique to identify specific structures known as key frames. The tests performed with a 224 by 224 image dimension showed an overall map of 77.22%. Additionally, mAPs of 82.16%, 69.1%, 66.5%, and 75.76% are attained for each of the four zones. The testing findings show that the vehicle re-identification method's accuracy varies significantly between various zones

Chatterjee et al. (2023) [3] Efficient DET-based architectures and Faster Region-Based Convolutional Neural Networks (Faster RCNN) for identifying human faces and arms. At the post-processing stage, a collective (stacked) technique has enhanced the detection ability to differentiate between human faces and firearms. In our ensemble (object) detection technique, a Network labeled with the ground truth is supplied to a detector's training architecture. An assembly approach called "weighted box fusion" combines the estimated bounding boxes from several item identification methods.

Dundar et al. (2022) [4] converted an image into a stream of position and appearance data and then reassembled the image from the factorized parts. The attitude representation should include a collection of dependable and closely spaced landmarks to make it easier to rebuild the input image. Species want our learned landmarks to concentrate on the foreground object of interest in an image from two disturbed variants: one where the appearance (color, lighting, texture) and one where the pose (position, perspective) of the object are disconcerted. The exploration described here investigates the effects of factoring the reconstruction task into separate foreground and background reconstructions, unsupervised, facilitating the model to be conditioned on only the foreground. Video input is divided into pre-accident, accident, and post-accident phases.

Gowri et al. (2023) [5], a combination of low-level data taken from certain video frames, as utilizing the entire video as input might introduce noise and redundancy into the learning operation. Data utilizes low-level characteristics and combines them through model training with adaptive enhancement, an ensemble learning approach, to solve the binary classification issue of violent video categorization. For hockey fights, movies, and audience violence, the suggested work archives outcomes were competitive with an accuracy of 90.62%, 91.70%, and 94.3%.

According to Guillermo et al. (2020), the primary data source should have been live CCTV footage; however, owing to experimental needs and limited access to resources, pertinent web films were obtained in their place. By dividing the frame rate by the duration of the video (in seconds), one may get the number of frames in a video. A physical assault video, for instance, lasts for 5 minutes and 15 seconds. The following annotations are training data sets. The XML documents should first be converted to a CSV file, followed by other procedures. Three datasets were executed after this neural network had been trained, and each one had a distinct function. One operates on the input of frames or photos, and the other on the video.

Holla et al. (2021) [6], their dataset is a multilayer dataset part submission, but the sliding window technique is the most direct and thorough check of the entire data in the Internet of Things. The user needs to look at all possible locations in the image, but also needs to look at different sizes. This is because models are often trained on a particular range. Huang et al. (2021), with the height and width that go along with the item's coordinates, this Technique produces the precise location of an object within an image. The characteristics of the algorithms mentioned above are used in this work. The detection technique provides the class name and the bounding box's x and y coordinates, width, and height.

Insaf Adjabi et al. (2021) [7], after preprocessing the coordinates by thresholding, the median filter was used to reduce noise while keeping face attractiveness and improving operational results. The number of basis components determines the length of the binary code string. Histograms of the binary codes can easily represent regions of an image for the pixels and other binary code-generating descriptors.

Kim et al. (2021) [8] A cutting-edge Subpixel Block Attention (SBA) module that recalibrates block-wise features to reduce block-wise discontinuities further to boost the computation-memory ratio for image restoration networks. According to an analysis of the computation-memory balance, the encoder-decoder construction is ideal for controlling. As explained in the introduction, the computational cost is a statistic that depends on the data. A basic network architecture is used to better understand the link between the computational cost and the number of parameters. A cutting-edge Subpixel Block Attention (SBA) module that adjusts block-wise characteristics, reduces block-wise discontinuities, and boosts performance.

Koh et al. (2016) [9]. Many businesses have recently installed network cameras for various uses, including watching traffic and taking in the scenery. Without a password, the data is accessible to anybody with an Internet connection. Even if the cameras are not being used for surveillance, by correctly integrating them into the existing surveillance systems, they may be used to improve public safety. Law enforcement may use CCTVs to expand coverage and monitor suspicious activity in real time. There are several difficulties when integrating public cameras into a surveillance system, including erroneous locations, inconsistent sources, and various approaches to getting visual data.

Lee et al. (2020) [10] experimented with the design of the automatic database creation method for local optimum is a revolutionary approach that automatically generates a high-quality training database. Preprocessing, backdrop modeling, object identification, and post-processing are the four components of the suggested process. A consideration is delivered to the preprocessing section, which does augmentation. A backdrop image is created from successive frames of an image. The object detection component uses a deep learning detector to extract object lists, which will be improved by local optimization.

Li et al. (2022) [11] introduced a unique logarithmic observer for online recovery of feature depth by employing the reduced-order observer concept. The given observer has several benefits over ordinary observers, including global convergence, a quicker rate of error structure convergence than exponential error structure, a less restrictive observability criterion, and increased resilience against noisy measurements. The usefulness of the given observer in realistic situations is further validated by actual tests using the Kinect v2 sensor.

Li et al. (2022). Gaussian distribution. Due to its sensitivity to the data collected, the One-Class Classifier performs the classification task. It views the features retrieved from single-encoded movies as the target class in double compression detection. The calculated decision boundary is erroneous if the probability density of a training set does not match that of a testing set. Using temporal correlation-based clustering, the training set was divided into eight subsets of data, from which eight matching one-class sub-classifiers were created. One class categorization, frequently used to find anomalies, simply requires the target class.

Li et al. (2020) [12], a deep learning-based technique for video image identification in the system to get multiple abstract representations of the input samples, the forward transformation is carried out by identifying the mathematical framework. Pre-training is carried out before using the deep learning method. Continuous training is used to accomplish this pre-training procedure using the limited Boltzmann machine as a model in between layers. The video emulation treats each pixel of the same surface configured in the video as a separate object, using streamlines to indicate the object's regional motion and its pixel motion.

Liu et al. (2021) [13]. Spatial-temporal data becomes ambiguous as a result of error propagation. A two-stream STMIF-Inter network, which fuses spatial, temporal, and multiscale data, is designed by the user for inter-coding. The modalities include intra/inter coding modes and partition modes. These two factors both have a significant influence on limiting the spread of errors. A shortcut convolution architecture is created for this network to learn the multiscale and multi-grained consideration of data that is connected with the partition. A two-stream convolution architecture is designed to understand the distortion propagation across frames and the spatial-temporal texture.

Mai Magdy et al. (2019) [14]. Moving cars is a possible risk; objects in this work should be identified for a safety study of people crossing a road. The findings of object Classification are employed in a Kalman filter because relationships between observed pedestrians and nearby cars must be considered. The most recent deep learning technologies for recognizing and classifying traffic users, such as automobiles and pedestrians, use a Kalman filter, and the objects are monitored while taking into account the class with the highest probability within the specified estimation error.

Morikawa. C et al. (2021) [15] Mobile image analysis and applications that use computer vision highlight these obstacles while clarifying how the algorithms have been modified or adapted to meet performance and accuracy requirements, even though images captured by cameras on mobile devices can be managed with generic image processing algorithms due to several constraints and outside issues. A valuable tool for academics wishing to apply image processing and computer vision techniques to mobile device-related real-world scenarios and operations.

Mounir. R et al. (2023) [16], LSTM forecasts of high-level characteristics calculated by a typical deep learning backbone, an attention mechanism filters the input features before the forecasting stage in spatial Segmentation using the object's stable description. The self-learned attention maps successfully focus on the item due to perceptual prediction. Additionally, users use lengthy, multi-day, 25 FPS recordings from ongoing wildlife video monitoring to illustrate our methodology.

Nie et al. (2021) [17] Step one involved performing an image de-noising Technique and visualizing occurrences as a series of event illustrations. High-performance Motion Vector Fields (MVF) between nearby de-noising event images with minimal computing go to step two of the extended 3D partial recursive search (E-3DPRS) approach. High-time resolution MVF represents the object's motion trajectory. Furthermore, the severely ill-posed DE-blurring procedure is lessened by using MVF as a priori knowledge to recover potentially crisp images. The third phase used MVFs and a crisp image sequence to rebuild a high-frame-rate, sharp video.

Oh et al. (2022) [18]. Data augmentation is a typical preprocessing technique that has been shown to decrease the amount of training data and increase its quantity. The dataset consists of training, validation, and testing sets, each randomly divided into three groups of eight. The inspectors needed to know the trip distance during the inspection to pinpoint the precise position of a flaw using the robot's referenced travel distance. This module, however, disregarded spatial attention, which is crucial in determining where in the image the model should concentrate.

Singh et al. (2018), used real-time video streams from surveillance cameras, decisions regarding violence are made. It gathers human activity video feeds and produces the aspects of violence by using motion tracking, which slices the video frames according to whether or not there are moving objects. It computes the optical flow to determine the violent flow descriptors for every pixel in the image. These descriptions of violent flow utilize several machine-learning methods to identify violent incidents.

In Thakare et al. (2022) [19], the next step is to define usual and accidental interactions using an iterative training process; these visualizations are often used to highlight the affected area. In conclusion, users develop high-level linguistic descriptions to assess the context and seriousness of an accident. However, video inputs from the CCTV camera may be used online to identify real-time road accident situations. Additionally, the model may be retrained for online surveillance with minimum supervision. Kun Jiang et al. (2023) recorded the number of identified vehicles in each lane and on each side. The Algorithm allows for the time for the green light based on the proportional traffic density and these facts. The concept of edges or lines indicating a departure from one is essential for detecting things in a photograph to compare a group of related pixels to another particular group.

Table 1. Summary of Problem Statement in Previous Research Work

Author	Technique Used	Data Set	Result	Drawbacks
R. Chatterjee et al. 2023[3]	Faster Region-Based Convolutional Neural Networks (Faster RCNN)	Internet Movie Firearms Database	Mean average precisions were 77.02%, 16.40%, and 29.73% for the mAP0.5, mAP0.75, and mAP [0.500.95], respectively.	In excess, complexity requires thorough, extensive data.
B. A. Holla et al 2023[6]	Vehicle Re-Identification	Zone-specific vehicle re-identification dataset	Map of 77.22%	Getting the Input source image is complex, especially during traffic speed
Yawen Pang et al 2023[20]	Self-organizing mapping neural network	Dancing video to Dancing image	Accuracy, 9.34%	Different patterns of images are Complex

	(SOM)			
Xiaodi Huang et al. 2023[21]	Improved ant colony algorithm	vehicle running capture Image	Image classification data	Sometimes, abnormal activities are not captured
Mounir et al. 2023[16]	Long-short-term memory cell (LSTM), Automated programming	Streaming Wildlife Video	Spatial and temporal Segmentation	The accuracy of segmented images is low
Kun Jiang et al. 2023[22]	Non-negative Analysis Representation (DSNAR)	EYaleB and Scene15 datasets	classification without complex	Efficiency is common in the max pooling layer and the convolution layer in the classification technique
C. Oh et al. 2022[18]	YOLOv5 architecture, Deep learning	CCTV video dataset	precision (mAP) of 75.9%	Sometimes, Face recognition images are not clear from CCTV video
Snehil G Jaiswal et al. 2021 [23]	Machine Learning Approach	Multi-Resolution Local Binary Patterns (MLBP) and Fuzzy Histogram of Optical Flow Orientations (FHOFH)	Accuracy of 90.62% on hockey fights, 91.70% on movies, and 94.3% on crowd violence.	Low data sets such as aspects such as features and classifiers
Li et al. 2022[11]	Subtractive pixel adjacency matrix (SPAM), Gaussian distribution.	One-Class, Video	Improved classification	High noise while preprocessing the image
Rimsha Rafique et al., 2022[24]	Deep fake detection and classification using error-level analysis and deep learning	Support Vector Machines and K-Nearest Neighbors	accuracy of 89.5%	Image object has low image detection

Tran et al. (2022) [25] Closed-Circuit Television (CCTV) recording is used to identify forests, and the process relies heavily on object detection. Each multiple process's object detection backbone was given rising values to classify the damaged regions. As an outcome, the issue might be characterized as a regression problem from the viewpoint of machine learning. When smoke or damage arises, AI-powered CCTV cameras quickly identify the problem and communicate the data to a database server.

Uzair et al. (2021) [26] Prior studies on tiny target identification have also employed visible spectrum photos and video sequences (RGB/grayscale). The majority of earlier studies on tiny target identification in the RGB image or video domain generally used artificial targets. Techniques for backdrop modeling that are both parametric and non-parametric. Select the well-known Mixture of the Gaussian model under the parametric heading. The adaptive filtering process in cells that sense light reduces clutter and broadens the range of potential input signal changes, improving the contrast of the target backdrop. The top-performing Algorithm's area under the Receiver Operating Characteristic (AUROC) curve has been enhanced by 75.4% just by bioprocessing.

3. MATERIALS AND METHODS

Initially, the 400-testing video sequence is fed into the outline, which creates frames at a rate of five frames per second (fps). Then, identical frames are removed using the key frame extraction approach. The next step in the data augmentation violent activity detection outline is to identify the specific object with two distinct modules in trajectory-based and non-object-centric methods. The video streaming input flow includes the detection of tables, etc., and tracking is the act of following a moving object. Image preprocessing is a basic step in computer vision and image processing. An Adafruit cloud sends the data to a data log, converting information from IFTTT to a mobile notification.

Fig. 2 shows the overall working outline; the recorded footage was examined to look for signs of anger and violence. The YOLO V7 algorithm comprises the Segmentation, feature extraction, classification, and preprocessing steps. The Inception-v3 classification method was used to identify the violent and unknown person in video frames and sound the warning. To construct each video's Mel spectral, the initial step was turning every video into an audio file. The audios without assault that lasted more than 40 seconds were divided in half in the second stage, making each entry into two separate entries in the dataset with an identical label.

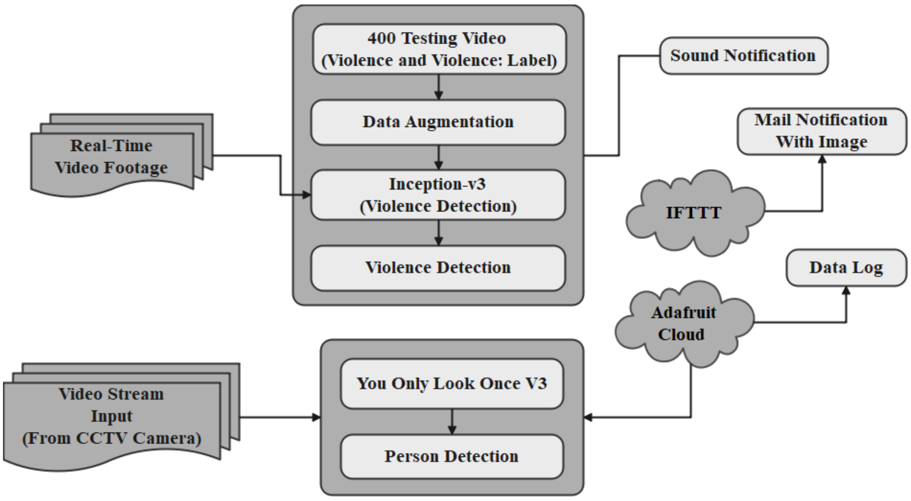


Fig. 2. Proposed Block Diagram

3.1. Dataset description

In videos, the Input source by the camera module gathers images as shown in Fig. 3, preprocessing reduces distortion and eliminates noise from the video streaming input, improving the quality of the concluding image. Preprocessing activities are often necessary before the major evaluation of data and data extraction, and they are usually referred to as radiometric or geometric adjustments.

Fig. 3. From the video Streaming output, A) Input Image, B) Preprocessing Image

The experiment used preprocessing techniques to convert RGB images to grayscale, downsize images, remove the background noise, and improve image quality. Consider the method is the process of editing a graphic so that the end outcome is considerably more suited for a certain purpose than the original image.

3.2. Data Augmentation Techniques

Data augmentation techniques extend the functionality of classic filters throughout the range of an image. Two pixels can be close to one another by occupying neighboring spatial locations or by having comparable values, potentially in a perceptually relevant way. Similarity relates to proximity in the range, whereas closeness refers to proximity in the domain. The image is cropped during the cropping process, which reduces the input size. Rotation of the image, ranging from 1 to 359 degrees. Moving the representation along the x- or y-axis (left, right, up, or down) is known as translation.

$$BF\ [I]_\rho = \frac{1}{W_p} \sum\nolimits_{q \in s} G_{\sigma s}\ (|\rho\ - q|) G_{\sigma r}(\left|I_q - I_q\right|) I_\rho \qquad (1)$$

By using a ρ that indicates nonlinear, q linear with a combination of neighboring pixel values, the Median filter smooths the image while maintaining the image's edges. It blends shades of grey or colors based on how similar their blur area is removed with another geometrically and visually. A number that represents the W_p weighted average of the pixels nearby is assigned to each pixel in the image. The fundamental $G_{\sigma r}$ is to provide better density to pixels closer to the current center pixel in both the spatial and range domains.

3.3. YOLO V3 Algorithm

YOLOv3 is a deep learning model for target detection, When a rectangular bounding box detects a single object, the image is used. The crowd employed FCN and OpenCV in conjunction with the YOLO V3 approach for object identification. To pre-train, consider identifying obstacles such as pits, barriers, speed bumps, and other obstacles. The category is at the top of the bounding box after people classify the image using SSD, using only one pass.

Arrange image: Pre-trained images serve as the basis for image prediction. Image locale: The bounding box is used to locate the object's location inside the image. Image Recognition: Following object recognition, a category is included. Fig. 4 shows computer vision; object detection is a method used to locate and identify things inside an image or video. The Technique of locating one or more items correctly using bounding boxes—rectangular rectangles that surround the objects—is known as image navigation.

$$IoU = \frac{Area\ of\ overlap}{area\ of\ union} \qquad (2)$$

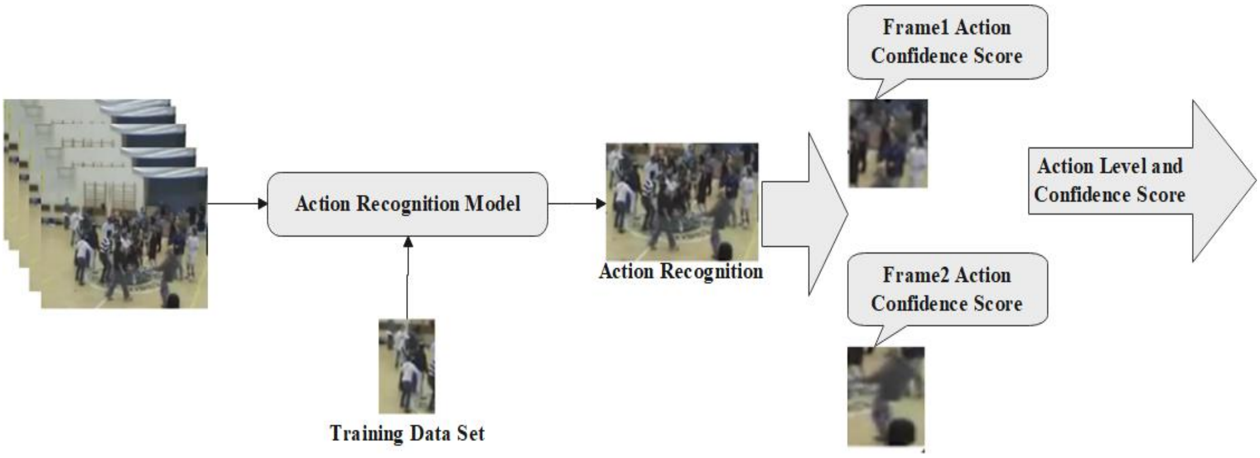

Frame1 Action Confidence Score
Action Recognition Model
Action Recognition
Action Level and Confidence Score
Frame2 Action Confidence Score
Training Data Set

Fig. 4. Working flow model of Inception-v3.

This procedure is occasionally mistaken with image recognition or classification, which attempts to identify which category or class an object or an item inside an image belongs to. The following graphic is a visual depiction of the explanation that came before it. "Person" is the object that was identified in the image.

3.3.1 Algorithm steps YOLO V3

Step 1: Classification of the Input image and extracting a classified output from the input image, an obstacle is run across it.

Step 2: Next, a sequence of fully connected layers that predict bounding box coordinates and class probabilities is applied to the features.

Step 3: A recognition model is created from the image, and the task assigned to each cell is to forecast a set of bounding boxes and class probabilities.

Step 4: A collection of frames and probability distributions for every movement of a human and edge up the confidence label.

3.4. Inception-v3

Very recently, a series of studies have demonstrated that the deep learning method of Convolutional Neural Networks (CNN) is very effective in representing violence detection[30,31,33], Inception-v3 is a convolutional neural network architecture from the Inception family that makes several improvements, including using label Smoothing. The transfer learning approach is applied in Inception-v3. Since we only applied five distinct mammals, the number of output nodes will change to five. Also, it should maintain the settings of the previous layer, delete the last layer, and then retrain the new final layer using the mammal dataset. The cross-entropy cost function is applied to calculate the error between the outcome of the Softmax layer and the label vector of the specified sample category to alter the weight parameter, and the last layer of the model is trained using the backpropagation approach.

However, gathering enough information to produce a useful dataset can sometimes be challenging. By adding synthetic data to the dataset, the data extension process is commonly used for producing a huge volume of training data. This false information might be entirely new data generated from the dataset's existing data, or it could be models of the dataset's current data with just slight modifications.

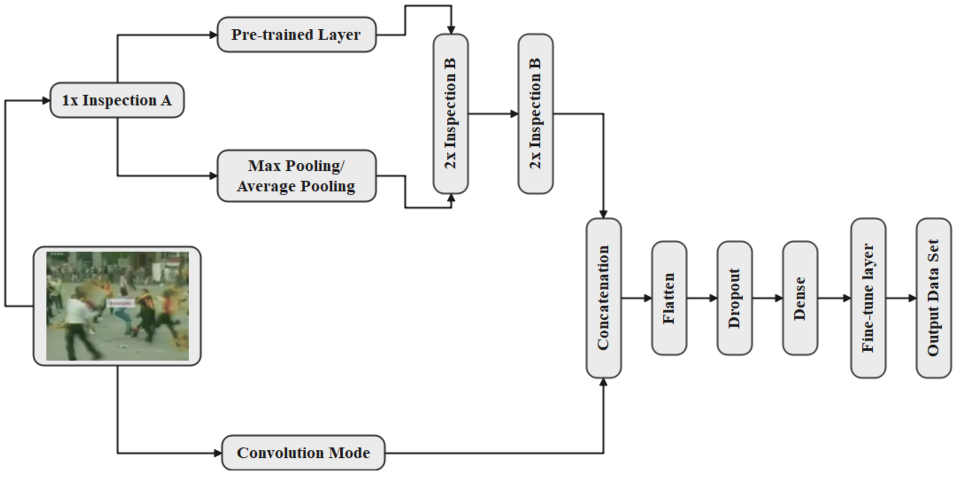


Fig. 6. Working flow model of Inception-v3

CCTV footage with a 256x256 image resolution was the image employed in this experiment. One single video containing fifty frames, each made up of the training set of multiple layers. Training and validation data are separated 90:10 in the training data set. The 2x Inspection B extracts features from images, and by employing the convolution kernel, the number of parameters may be significantly decreased.

$$softmax(x_i) = \frac{\exp(x_i)}{\sum_i(\exp(x_j)} \qquad (3)$$

4x Inspection C refers to using the same convolution kernel's parameters across the image, with the convolution kernel's weight being constant despite the image's various placements. 8x Inspection D number of factors in the convolution kernel may be significantly decreased by using weight sharing in the fine-tuning procedure.

$$loss\ (x, class) = weight\ [class]\ (-x[class] + \log(\textstyle\sum_j \exp(x[j]))) \qquad (4)$$

Fig. 6 shows that each classifier's layer is transformed into a probability distribution using the soft-max function. Every output element is contained inside the period, and the total of all the output parts is 1. Fine-tune layer with 3x injection with often positioned after the convolution layer employed to lower the characteristic dimension of the output of the density layer. The weight that every class is given minority classes more weights in order to lessen the impact of class imbalance mini-batch; the losses are averaged over observations.

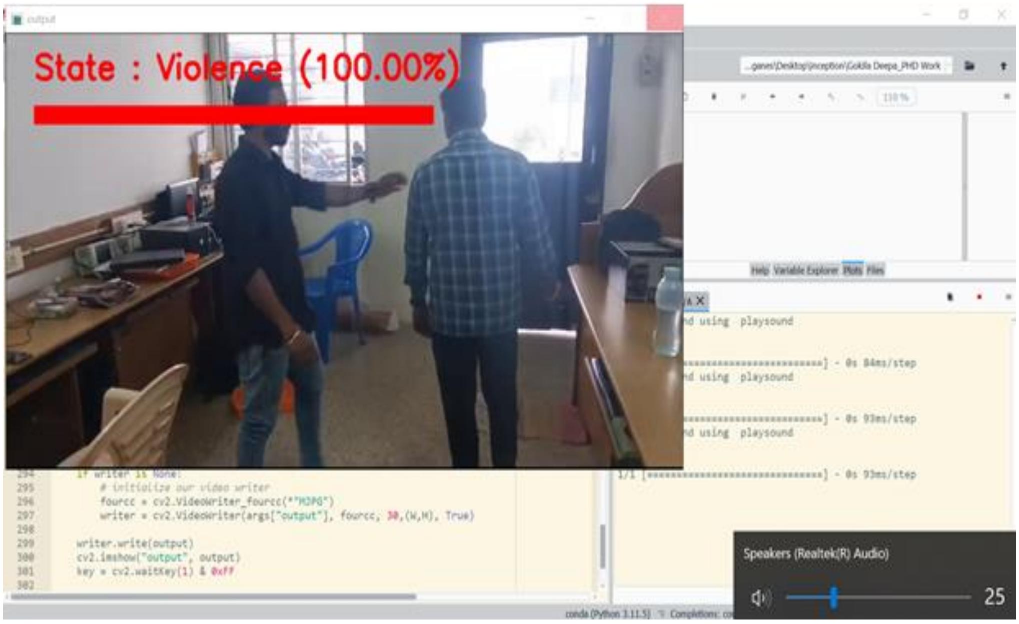
4. RESULTS AND DISCUSSION

In the experimental calculation, violence detection footage can be converted into image augmentation and used to classify an image with better accuracy. Modules: This system is split into five sections based on the roles of each image processing technique. Data training and testing of each video data is the focus of this module. Data Augmentation and Data Enhancement are its two submodules. Data augmentation is a technique for adding to the existing data; the main purpose of augmentation is to expand the amount of data that is obtainable.

Table 2: Input Parameters

Simulation Limits	Values
Number of Training Video	400
Software	Anaconda
Programing Language	Python

4.1 Violence detection


output
State : Violence (100.00%)
...ganes\Desktop\inception\Gokila Deepa_PHD Work
110 %
Help Variable Explorer Plots Files
Speakers (Realtek(R) Audio)
25
conda (Python 3.11.5)
Completions: co

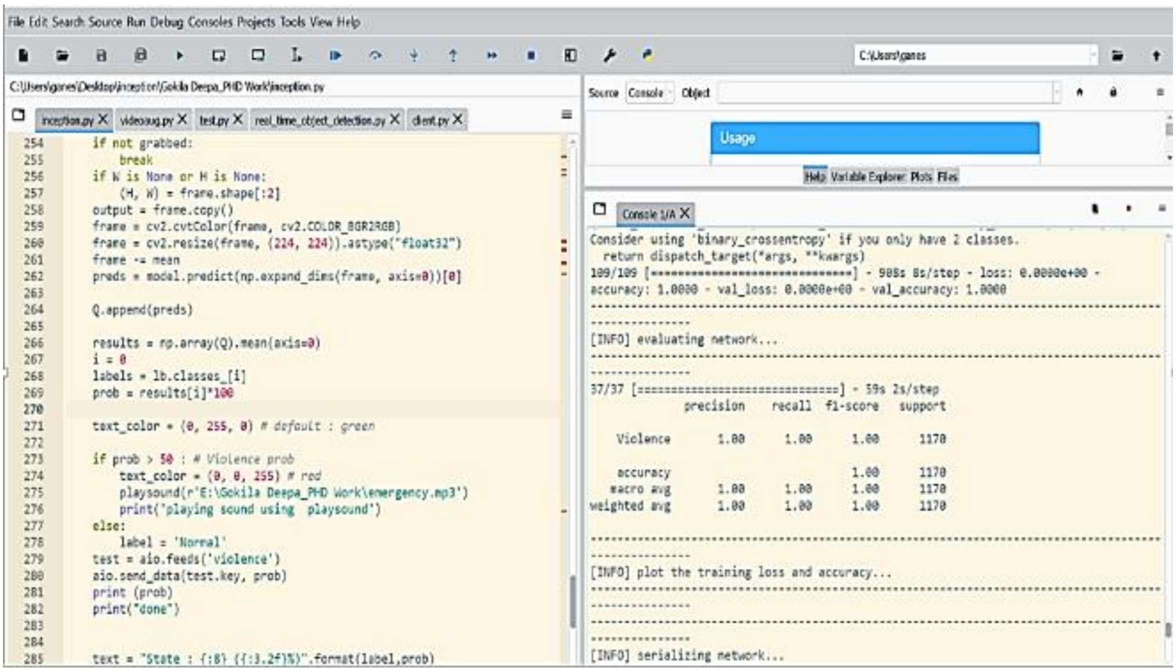
```
    if writer is None:
        # initialize our video writer
        fourcc = cv2.VideoWriter_fourcc(*"MJPG")
        writer = cv2.VideoWriter(args["output"], fourcc, 30,(W,H), True)

    writer.write(output)
    cv2.imshow("output", output)
    key = cv2.waitKey(1) & 0xff
```

```
nd using  playsound

=========================] - 0s 84ms/step
nd using  playsound

=========================] - 0s 93ms/step
nd using  playsound

1/1 [==============================] - 0s 93ms/step
```

Fig. 7. Violence detection Output (Testing data set)

Fig. 7 shows that video live detection is implemented, the input source is streaming video from the webcam, or the source camera is transformed into video data converted into frames, and the concept behind the database is image processing. In other words, the anticipated video is submitted, yielding findings about either non-human or human aggression, and the objects detected are then kept in a database.

Fig. 8. Shows the output of data collected from violence detection-based movements and actions that are identified as normal or abnormal activities. The data calculated based on accuracy recall were current in the database during either a human-violence or a non-human-violence situation.

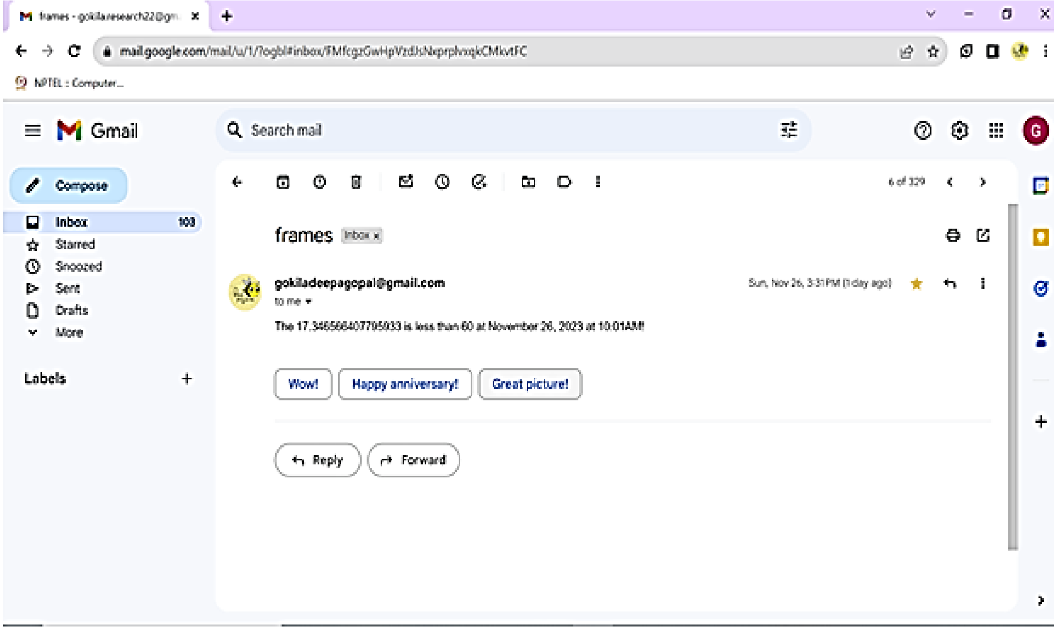

File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\ganes\Desktop\inception\Gokila Deepa_PHD Work\inception.py
inception.py
videoaug.py
test.py
real_time_object_detection.py
client.py
C:\Users\ganes
Source
Console
Object
Usage
Help
Variable Explorer
Plots
Files
Console 1/A

```
    if not grabbed:
        break
    if W is None or H is None:
        (H, W) = frame.shape[:2]
    output = frame.copy()
    frame = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)
    frame = cv2.resize(frame, (224, 224)).astype("float32")
    frame -= mean
    preds = model.predict(np.expand_dims(frame, axis=0))[0]

    Q.append(preds)

    results = np.array(Q).mean(axis=0)
    i = 0
    labels = lb.classes_[i]
    prob = results[i]*100

    text_color = (0, 255, 0) # default : green

    if prob > 50 : # Violence prob
        text_color = (0, 0, 255) # red
        playsound(r'E:\Gokila Deepa_PHD Work\emergency.mp3')
        print('playing sound using  playsound')
    else:
        label = 'Normal'
    test = aio.feeds('violence')
    aio.send_data(test.key, prob)
    print (prob)
    print("done")


    text = "State : {:8} ({:3.2f}%)".format(label,prob)
```

```
Consider using 'binary_crossentropy' if you only have 2 classes.
  return dispatch_target(*args, **kwargs)
109/109 [==============================] - 908s 8s/step - loss: 0.0000e+00 -
accuracy: 1.0000 - val_loss: 0.0000e+00 - val_accuracy: 1.0000
--------------------------------------------------------------------------------
---------------
[INFO] evaluating network...
--------------------------------------------------------------------------------
---------------
37/37 [==============================] - 59s 2s/step
              precision    recall  f1-score   support

    Violence       1.00      1.00      1.00      1170

    accuracy                           1.00      1170
   macro avg       1.00      1.00      1.00      1170
weighted avg       1.00      1.00      1.00      1170

--------------------------------------------------------------------------------
---------------
[INFO] plot the training loss and accuracy...
--------------------------------------------------------------------------------
---------------
--------------------------------------------------------------------------------
---------------
[INFO] serializing network...
```

Fig. 8. Output of the data set from the violence detection output

4.2 *Ada-fruit Mobile notification with image*



frames Inbox

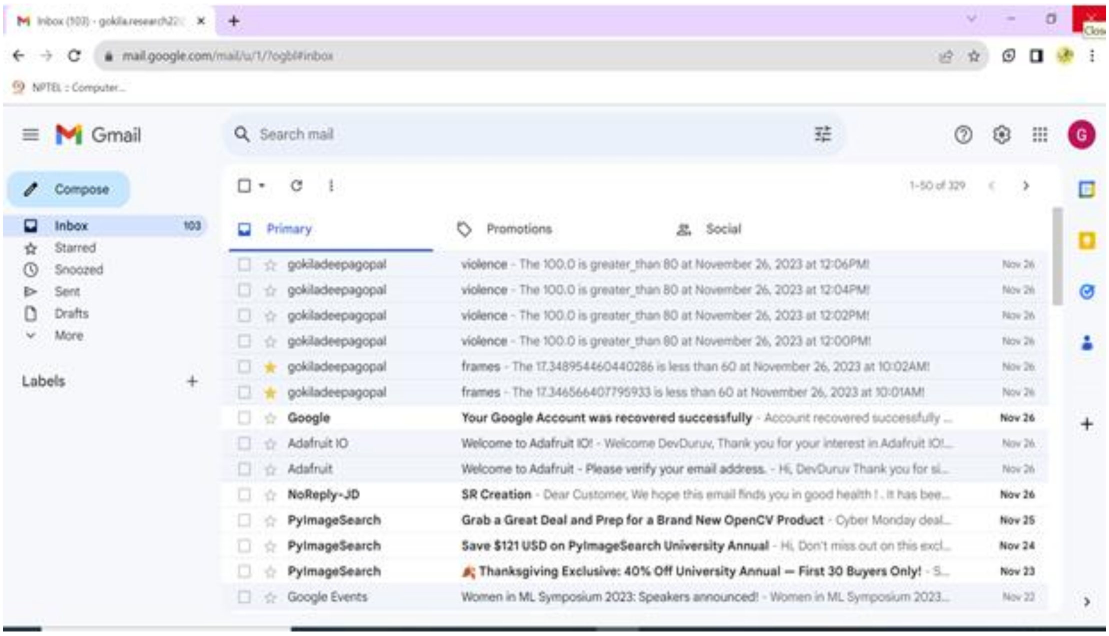
gokiladeepagopal@gmail.com
to me
Sun, Nov 26, 3:31PM (1 day ago)

The 17.346566407795033 is less than 60 at November 26, 2023 at 10:01AM!

Fig. 9. Training Dataset 1 Email Notification

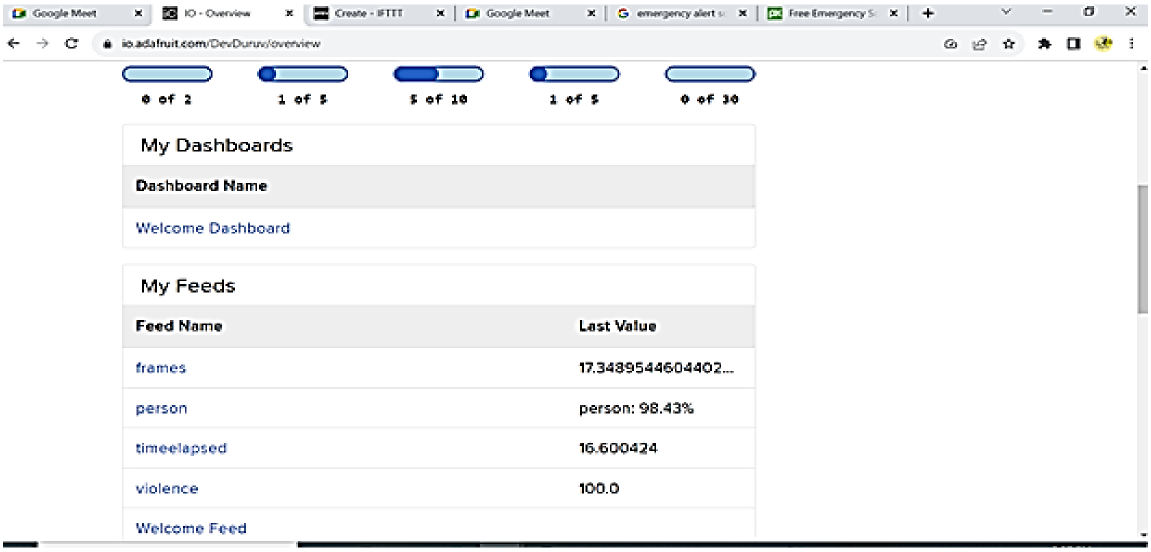
Fig. 9. shows the real-time notification whether the data rate is normal or abnormal, does not match the database, the individual gets identified as a stranger, and a warning to the user is sent along with an audio output to warn and scare an intruder. After completing all of the preceding actions, the main.py file was executed. The security camera continually scans its field of view for any movement. When it does, an email alert is sent to the user's email address along with the image that was continuously mail detected using Fig. 10.

Fig. 10. Classification is the process of classifying remotely sensed data with the user's input and guidance through training data. Automatic identification of natural groups or structures within a remote sensing data set is known as unsupervised classification.



gokiladeepagopal	violence - The 100.0 is greater_than 80 at November 26, 2023 at 12:06PM!	Nov 26
gokiladeepagopal	violence - The 100.0 is greater_than 80 at November 26, 2023 at 12:04PM!	Nov 26
gokiladeepagopal	violence - The 100.0 is greater_than 80 at November 26, 2023 at 12:02PM!	Nov 26
gokiladeepagopal	violence - The 100.0 is greater_than 80 at November 26, 2023 at 12:00PM!	Nov 26
gokiladeepagopal	frames - The 17.348954460440286 is less than 60 at November 26, 2023 at 10:02AM!	Nov 26
gokiladeepagopal	frames - The 17.346566407795933 is less than 60 at November 26, 2023 at 10:01AM!	Nov 26
Google	**Your Google Account was recovered successfully** - Account recovered successfully ...	Nov 26
Adafruit IO	Welcome to Adafruit IO! - Welcome DevDuruv, Thank you for your interest in Adafruit IO!...	Nov 26
Adafruit	Welcome to Adafruit - Please verify your email address. - Hi, DevDuruv Thank you for si...	Nov 26
NoReply-JD	**SR Creation** - Dear Customer, We hope this email finds you in good health ! . It has bee...	Nov 26
PyImageSearch	**Grab a Great Deal and Prep for a Brand New OpenCV Product** - Cyber Monday deal...	Nov 25
PyImageSearch	**Save $121 USD on PyImageSearch University Annual** - Hi, Don't miss out on this excl...	Nov 24
PyImageSearch	**Thanksgiving Exclusive: 40% Off University Annual — First 30 Buyers Only!** - S...	Nov 23
Google Events	Women in ML Symposium 2023: Speakers announced! - Women in ML Symposium 2023...	Nov 22

Fig. 10. Training Dataset 2: Continuous Email Notification



My Dashboards

Dashboard Name
Welcome Dashboard

My Feeds

Feed Name	Last Value
frames	17.3489544604402...
person	person: 98.43%
timeelapsed	16.600424
violence	100.0
Welcome Feed	

Fig. 11. Training Dataset 2: Dashboard

Fig. 11 shows a message delivered to the user through a module informing them that someone they know has returned home if the image matches the database and the individual is a customer. For a certain set of test Image Storage, the calculation is based on a confusion matrix that may be employed to assess how well the classification models performed, and output-specified identification can be established if the test data's real values are available.

4.3 Comparison Analysis of Classification Technique

Fig. 12 shows that the analysis of accuracy performance is the proposed Inception-v3 algorithm. The existing Algorithm is K-Nearest Neighbors (KNN) 85.00 % [26], Weighted Average on Linear SVM 96.98 %. The proposed V3 pretrained network algorithm method is based on a performance level of 100% while analyzing the different steps using a performance level in accuracy, sensitivity, and specificity of CCTV video output.

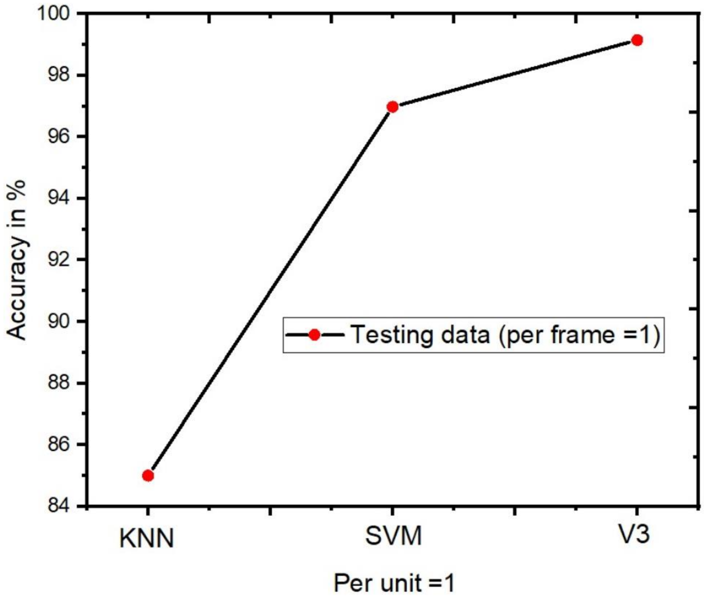


Fig. 12. Compression analysis of Recall output

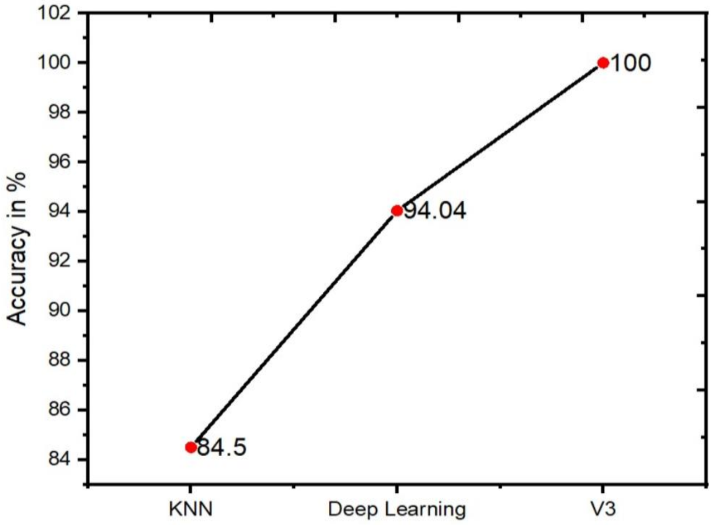


Fig. 13. Compression analysis of Accuracy output

Accuracy = (TP + TN) / (P + N) (5)

Ex: (50+50)/ (50+50+2+2) = 0.99245 x 100 = 100 %

Fig. 13 shows that the analysis of accuracy performance is the proposed Inception-v3 algorithm. The existing Algorithm is Computer vision Violent video Classification using Deep learning 94.09% [19], Three- Residual Network and K-Nearest Neighbor [24] 84.5 %; the proposed V3 pre-trained network algorithm method-based performance level of 100% while analyzing the different steps using and performance level in accuracy, sensitivity, specificity CCTV video output.

5. CONCLUSION

Violence detection from input data set videos is analyzed for human behavior in several activities. However, the majority of individuals train and test video monitoring to be object detection on their data, each containing different irregular detection videos that are converted into several frames. Combining several video analyses based on violent and non-violent events separated into training, validation, and testing sets, the user develops a model for detecting violence. The outcome of real-time detection uses and abuses of video recordings, particularly with the usage of image processing technology. YOLO V3 is a sophisticated object recognition method that can identify several objects, like a table, chair, etc. The Inception-v3 overcame the detection of violence; the testing improves the accuracy rate and gating in real-time information of mobile, and gains an accuracy is 99.15%. As our experimental results indicated, the proposed method modernizes accurate results and identifies real-time violence detection.

5 FUTURE WORK

In the future, classify violence detection using remote sensing techniques and advanced AI (Artificial Intelligence) networking techniques. And also, using a pre-trained ImageNet approach with a neural network, it extracts frame-level characteristics from a video. The frame level's features are then combined using a short-term long-term memory variable that employs leaky rectified linear units and completely linked layers. Convolutional neural networks can capture localized spatiotemporal information in combination with lengthy short-term memory, which allows for the analysis of each movement in videos.

REFERENCES

[1] Rao, M. Koteswara, and P. M. Ashok Kumar. "Multi-level glowworm swarm convolution neural networks for abnormal event detection in online surveillance video." International Journal of Information Technology (2024): 1-9.

[2] Bhatti M. T, M. G. Khan, M. Aslam, and M. J. Fiaz, “Weapon detection in real-time CCTV videos using deep learning”, IEEE Access, vol. 9, pp. 34366-34382, 2021.

[3] Chatterjee. R, A. Chatterjee, M. R. Pradhan, B. Acharya, and T. Choudhury, "A Deep learning-based efficient firearms monitoring technique for building secure smart cities," IEEE Access, vol. 11, pp. 37515-37524, 2023.

[4] Dundar, K. J. Shih, A. Garg, R. Pottorf, A. Tao, and B. Catanzaro, “Unsupervised Disentanglement of Pose, appearance, and Background from Images and Videos," IEEE Transactions on Pattern Analysis and machine intelligence, vol. 44, no. 7, pp. 3883-3894, 2022.

[5] Gowri. D, S. Shiva Prasad, O. Sai Kiran, Mr. M. Mysaiah, "Image processing-based fire detection by using raspberry PI," Journal of Engineering Sciences, vol. 14 Issue. 06, 2023.

[6] Holla B. A, M. M. M. Pai, U. Verma, and R. M. Pai, “Enhanced vehicle re-identification for smart city applications using zone-specific surveillance”, IEEE Access, vol. 11, pp. 29234-29249, 2023.

[7] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, Sébastien Jacques, Multi-Block Color-Binarized, “Statistical images for single-sample face recognition”, Sensors, 2021.

[8] Kim. T, C. Shin, S. Lee, and S. Lee, “Block-attentive subpixel prediction networks for computationally efficient image restoration”, IEEE Access, vol. 9, pp. 90881-90895, 2021.

[9] Koh. Y et al., “Improve safety using public network cameras”, IEEE Symposium on Technologies for Homeland Security (HST), Waltham, MA, USA, pp. 1-5, 2016.

[10] Lee J. G. and J. W. Baek, “An automatic database generation algorithm for local optimization of CNN object detector for edge devices”, IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), Seoul, Korea (South), pp. 1-3, 2020.

[11] Li. X, H. Zhao, and H. Ding, “Logarithmic observation of feature depth for image-based visual serving”, IEEE Transactions on Automation Science and Engineering, vol. 19, no. 4, pp. 3549-3560, 2022.

[12] Li. A, B. Zheng, L. Li and C. Zhang, “Optical flow estimation and DE noising of video images based on deep learning models”, IEEE Access, vol. 8, pp. 144122-144135, 2020.

[13] Liu. X, K. Shi, Z. Wang, and J. Chen, “Exploit camera raw data for video super-resolution via hidden Markov model inference”, IEEE Transactions on image processing, vol. 30, pp. 2127-2140, 2021.

[14] Mai Magdy, Mohamed Waleed Fakhr, Fahima A. Maghraby, “Violence 4D: Violence detection in surveillance using 4D convolutional neural networks”, the institute of Engineering and technology, wile, pp 282-284, 2022.

[15] Morikawa, C., Kobayashi, M., Satoh, M, “Image and video processing on mobile devices: a survey”, The Visual Computer, no. 37, pp. 2931–2949, 2021.

[16] Mounir, R., Shahabaz, A., Gula, R, “Towards automated ethogramming: Cognitively-inspired event segmentation for streaming wildlife video monitoring, International Journal of computer vision, no. 131, pp. 2267–2297, 2023.

[17] Nie. K, X. Shi, S. Cheng, Z. Gao, and J. Xu, “High frame rate video reconstruction and DE-blurring based on dynamic and active pixel vision image sensor”, IEEE Transactions on Circuits and Systems for video technology, vol. 31, no. 8, pp. 2938-2952, 2021.

[18] Oh. C, L. M. Dang, D. Han, and H. Moon, “Robust sewer defect detection with text analysis based on deep learning”, IEEE Access, vol. 10, pp. 46224-46237, 2022.

[19] Thakare K. V, D. P. Dogra, H. Choi, H. Kim and I. -J. Kim, “Object interaction-based localization and description of road accident events using deep learning”, IEEE Transactions on intelligent transportation systems, vol. 23, no. 11, pp. 20601-20613, 2022.

[20] Yawen Pang and Yi Niu, “Dance video motion recognition based on computer vision and image processing”, Applied Artificial Intelligence, no. 37:1, 2023.

[21] Xiaodi Huang, Po Yun, Shuhui Wu, and Zhongfeng Hu, “Abnormal driving behavior detection based on an improved ant colony algorithm”, applied artificial intelligence, no. 37:1, 2023.

[22] Kun Jiang, Lei Zhu, and Qindong Sun, “Joint dual-structural constrained and non-negative analysis representation learning for pattern classification”, Applied Artificial Intelligence, 37:1, 2023.

[23] Jaiswal, Snehil G., and Sharad W. Mohod. "Implementation of Violence Detection System using Soft Computing Approach." Data Analytics and Management: Proceedings of ICDAM. Springer Singapore, 2021.

[24] Rafique, Rimsha, et al. "Deep fake detection and classification using error-level analysis and deep learning." Scientific reports 13.1 (2023): 7422.

[25] Tran D. Q, M. Park, Y. Jeon, J. Bak, and S. Park, “Forest-fire response system using deep-learning-based approaches with CCTV images and weather data”, IEEE Access, vol. 10, pp. 66061-66071, 2022.

[26] Uzair. M, R. S. Brinkworth and A. Finn, "Bio-inspired video enhancement for small moving target detection”, IEEE transactions on image processing, vol. 30, pp. 1232-1244, 2021.