Developing A Framework For Smart Surveillance System Using Machine Learning

Akhilandeswari nese bandhike[1] **, Dr. Venkataraju kallipalli**[2]
[1]Koneru Lakshmaiah Education Foundation, Department of Computer Science and Engineering, Vaddeswaram, Guntur, 522302, AP-India
[1]akhila30032@gmail.com

Abstract
There have recently been more rapid advances in AI and deep learning. Some of these advances provide intelligent surveillance systems autonomously capable of monitoring and analyzing data in most cases in real time. In this paper, a Smart Surveillance System is now described employing several deep-learning-based smart algorithms, specifically InceptionNet and Gated Recurrent Units (GRUs), for very high accuracy in detecting and classifying suspicious activities. The architecture of the system consists of frontend and backend modules. The backend modules consist of dataset acquisition, splitting, preprocessing, and training using the smart algorithms mentioned above. The system's user interface is provided in the front for registration, login, and data input. The security authorization mechanisms will ensure that only entries from classified authorized personnel can access the system. After authentication, the input data are sent to the training model to generate pertinent insights. The results will be presented to the user; thus, timely decision-making will be reinforced in the security monitoring. It is the adopted integrated system, which promises to offer the most reliable, scalable, and efficient modern surveillance applications in vulnerable settings like public places, transport terminals, and private establishments.

Keywords: *Deep Learning, CNN, GRU, InceptionNet, Security, Real-time Monitoring, Computer Vision.*

1. INTRODUCTION

All these changes have commanded the attention of cities across the world over a couple of years, which calls for the demand for intelligent and automated surveillance systems due to increased safety measures, crime prevention, and effective monitoring of critical facilities. Traditional systems, heavily relying on manual observation through closed-circuit television (CCTV) networks, are highly subject to human error and fatigue. Besides, inefficiency in handling vast amounts of video data in real-time is part of the challenges resulting from depending solely on human intervention. The fact that most city spaces expand as security challenges evolve has called for advanced technology that can facilitate proactive real-time threat detection and response. This subsequently led to a deep learning integration in surveillance and how visual data was interpreted and analyzed.

Deep learning is one of the very good promises of artificial intelligence used among other branches in the area of computer vision. This includes image classification, object detection, and video analytics. Such models as, InceptionNet, Gated Recurrent Units (GRU), have shown an impressive ability of the model to extract complex features and identify them from large-scale visual datasets. One good example of approach using these models into an intelligent surveillance framework is for modeling a detection system not just capable of detecting an object, but even understanding the contextual input, tracking movement behavior, and anticipating potentially suspicious action.

The complete smart surveillance system designed can swallow all these elements as it is built by integrating InceptionNet, and GRU to develop a backend model that has high accuracy in processing input video data. It is a modular architecture system which has two components-it's backend, which is responsible for training and inference using deep learning models, and the frontend, which acts as the user interface for the registration, authentication, data input, and result visualization. This bifurcation improves the robustness, maintainability, and scalability of the efficiency of real-time environments for data processing and user interaction.

The backend pipeline is ipso facto conditioned by the structure of its well-organized workflow model of the different operational units including loading the dataset; partitioning into training and testing data; noise reduction and other normalizing preprocessing techniques narrowly less advanced neural architectures for model training. This is where multi-scale features are made use of, with the InceptionNet model, in detecting the different objects at varying spatial resolutions recognized by CNN. For spatial pattern recognition as applied to identifying human poses, objects, or unusual activities in video frames; and GRU, a recurrent neural network based type, can loop all these into analyzing actions over time instead of isolated frames. In the meantime, as a frontend module, it is being developed to offer an interface that is secure yet user-friendly for end-users, such as surveillance operators or administrative personnel. It allows new users to register and existing users to log in using a credential-based authentication module. Once authenticated, users will have the capability of inputting data (for example, video streams or pictures) into the system. The backend will process this input and subsequently return such data in a meaningful way. Defined is the logout process where there is session management and data security, which is highly recommended for sensitive applications like surveillance in airports, government buildings, and corporate offices. While most traditional systems perform mere recording and storage, the design system intends to consider video inputs and perform some analysis on it as they come in order to afford real-time responses to threats. An example of such a reaction may be sending an alert to the operator on the occurrence of unauthorized access, unattended objects for some time, or anomalous behavior patterns. Such capabilities would end up adding greatly to situational awareness and decrease response time during emergencies. Moreover, the proposed modularity of the framework makes it convertible to diverse datasets and deployment environments. The models would need fine-tuning or complete retraining when new data is now available, so the system becomes updated not just with the times, but also with evolving security challenges. Moreover, deep learning does not require significant manual engineering of features, and hence the system can learn directly from raw data and generalizes better as the environment varies. Therefore, the application of deep learning models into an intelligent surveillance framework is a revolutionary landmark in the field of security and monitoring. With the combination of CNN, GRU, and InceptionNet architectures, all organized within a structured backend, and by further supporting the system through a securely intuitive frontend, the system acts as a massively well-equipped intelligent surveillance object. It bridged the gap that conventional systems had by automating, analyzing in real-time, and carrying the prediction of future occurrences on the system, thereby putting it in the line for next-generation surveillance frameworks.

This guide provides details to assist authors in preparing a paper for publication in JATIT so that there is a consistency among papers. These instructions give guidance on layout, style, illustrations and references and serve as a model for authors to emulate. Please follow these specifications closely as papers which do not meet the standards laid down, will not be published.

2. Related Work

Video surveillance, nowadays a large part of any modern security system, acts destinative in surveillance of public and private places to keep them safe. Many researchers have tried their luck to enhance the video surveillance systems through different learning techniques such as deep learning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and IOT. Below are some important contributions in the field concerning advancements in violence detection, anomaly detection, and surveillance optimization. Arshad et al. (2024) propose a smart surveillance system using CNNs which analyzes videos for events such as fire, abnormal activities, smart parking systems as well as detecting burglary. This approach works excellently to the limitations of after-investigation procedures by facilitating real-time detection, reducing all possible human effort in surveillance [1]. They further elaborated that in this perspective, Patel and Patel (2021) indicated how conventional CCTV cameras are enhanced by deep learning models and IoT, all using Raspberry Pi in their project for smart surveillance systems. Their system design detects fire and weapon, face masks on faces, and very significantly reacts to real-time situations like during the COVID-19 pandemic [2].

Tiwari et al. (2023) detailed an automated violence detection model using hybrid CNN-LSTM architecture, achieving 98.63% accuracy. The proposed model establishes robustness in violence identification in video footage by demonstrating its competence by combining the features of CNNs with those of LSTMs for temporal sequence learning [3]. The same vein was explored by Akole et al. (2023) in which a real-time violent activity detection system was proposed using MobileNetV2 and LSTM. Their system achieved a success rate of 94%, efficiently classified real-time video streams for violent activity detection [4].

Further advancement in real-time violence detection has been done by Siddiqui et al. (2023) in which YOLO (You Only Look Once) was incorporated to give the object detection feature to the violence detection system they developed. The system also detects weapons in violent situations and thus triggers alarms for immediate intervention by security [5]. Akole et al. (2023) also build on MobileNetV2, which is characterized by very low latency and great accuracy, and uses it together with LSTM for detecting violence in videos. The proposed model has performed well on tailor-made datasets that are composed of violence videos sourced from social media like YouTube and is an ideal resource with rapid response to security threats [6]. Ullah et al. (2022) proposed a framework for violence detection in Industrial IoT-based surveillance networks using a lightweight CNN model for object detection and ConvLSTM for video analysis. The framework enhances real-time detection but is a major reduction in the computational load, which suits resource-constrained IoT environments for such applications. The proposed framework thus improves over the classical means by 3.9% and thus is an efficient endeavor for industrial surveillance [7]. Discussing the anomaly detection in CCTV footage, Khanam and Roopa (2025) reported a deep learning model, which is based on MobileNetV2 and Bi-LSTM, providing performance of 94.43% in accuracy under varying illumination conditions as compared to best available models [8].

Following on this line, Marwaha et al. (2023) presented a problem challenge that needs to address developing real-time CCTV camera surveillance systems that will be able to analyze several terabytes of video. Smart surveillance systems have generally incorporated the use of machine learning along with image processing techniques to detect activities taking place in such public areas in real time and alert the local authorities to prevent violent incidents from occurring [9]. Also, Ramya et al. (2023) proposed an EfficientDet-based weapon detection system, mainly concerning real-time security surveillance. The system proved to be of high precision and accuracy in detecting knives and pistols, as a requirement towards public safety in sensitive environments [10].

Appavu and Babu (2023) sought further to demonstrate the effectiveness of even more CNNs in real-time violence detection while implementing their real-time violence detection system that was based on the Xception model. Their system applied the combination of CNN feature extraction and LSTM to interpret temporal sequences, achieving an impressive level of accuracy in violent event detection. Real-app app integration with that of authorities led to instant reporting, hence proving its feasibility during emergencies [11]. Meanwhile, Jain et al. (2023) developed a violence detection model consisting of a U-Net combined with a MobileNetV2 for spatial feature extraction, which uses an LSTM for temporal analysis. Their approach achieved a performance of 94% while exploiting a real-life dataset and an effective resource utilization strategy [12]. With regard to crowd violence detection, Gkountakos et al. (2021) proposed a sociotechnical architecture based on 3D CNNs that encompasses the processing of video footage coming from several sources: the CCTV cameras and body-worn cameras. In applying deep neural networks to real-time analysis, the system detects footage related to violent happenings in crowded environments. It has been tested on the Violent Flows dataset, attesting to its boastful attributes concerning crowd violence detection [13]. Similarly, Ditsanthia et al. (2018) were concerned with violence detection in video representation learning in a deep CNN and LSTM context. Among their main contributions was the introduction of multiscale convolutional features as a concern with changes in video data and hence an improvement in detection accuracy in the bad-cinema environment [14].

Following the works by Kumar et al., in 2024, on the MobileNetV2 and BiLSTM-based model, a remarkable accuracy of 98% can be claimed. His model has proved to efficiently detect violent acts from many types of raw videos, enhancing surveillance and law enforcement activities. This speaks of the promise of lightweight CNN architectures for real-time tasks in violence detection [15]. Aggarwal et al. (2024) further fine-tuned the method using MobileNetV2 and BiLSTM for violence detection on CCTV footage, reporting 96% accuracy. This method demonstrated the working capability of hybrid deep learning models for violence detection in myriad real-life situations [16].

The above studies testify to how much forward the surveillance systems have moved by highlighting the growing importance of deep learning and hybrid models in enhancing video analysis. The synergistic path along with CNNs, LSTMs, and other deep learning models just goes on to ensure the further accuracy, speed, and reliability of violence detection within real-time video footage so that the guarantee of security isched towards public safety. The future of intelligent surveillance systems will heavily rely on the progress of the systems in cities, industrial corridors, and public spaces.

3. System Design

The recommended smart surveillance system is comprised essentially of two modules: the backend processing engine and the frontend user interface. These modules are purposely fashioned to interact in detection and monitoring of potential threats on a real-time basis. This modular architecture enhances flexibility, maintainability, and separation of concerns with the possibility of independently developing, testing, and deploying the respective components.

The backend module takes care of all machine-learning and deep-learning functions. The operations begin by the loading of the curated dataset(s), consisting of surveillance images or sequences of video, most of which are then split into training and test sets. Preprocessing is carried out before training of models-batching by group resizing, normalization, filtering for noise, etc.-in order to enhance the reliability of feature extraction. The system employs a hybrid deep-learning architecture comprising InceptionNet, GRU, and CNN. InceptionNet captures multi-scale features, CNNs are used for spatial feature extraction, and GRUs model the temporal dynamics in the video streams. The training of these models is designed to identify suspicious behaviors or anomalous activity in the visual data and to assign the input with predictions of what the input is in real-time.

The front end acts as an interface for the users and encompasses functionalities like user registration, secure login, and submission of requests for processing. After the user is granted access, the user can upload or stream surveillance data; in turn, these inputs are forwarded to the backend for further assessment. The visualized outputs for any detected activities are relayed in real-time through the interface. The system has session management capabilities like logout functionality that promote security and integrity in its operation.

In its entire consideration, system design turns into a rich interactive space for intuitive human-machine interaction and deep-learning-intensive back-end deployment for smart cities, airports, or any area in high security.

4. Dataset

The Surveillance Camera Violent Dataset (SCVD), with its origins in Kaggle, provides the ground for developing and testing an intelligent surveillance system aimed at classifying normal and abnormal activities in surveillance videos. It works perfectly for a dataset that propagates deep learning models to classify activities that can potentially endanger human beings in different settings, like outdoor streets or indoor ambiance. A strong member of the SCVD is a collection of real-world surveillance video clips having diverse backgrounds, lighting, and motion dynamics, which serves as a good resource for training models against complex real-world scenarios.

4.1 Data Set Analysis:

The dataset is fairly balanced across its three categories: Normal, Violent, and Weaponized, having 872, 970, and 832 video clips respectively. This almost equal distribution guarantees training of deep learning models on enough samples from each class, thereby reducing the bias towards any one specific class. The slight higher count of Violence clips (970) also makes the model more receptive toward aggressive human behaviors which are important in real time threat detection. To adjust, however, the training and validation should take care of maintaining the class balance to be able to avoid producing skewed predictions mainly during real-world implementations as regards imbalances. The composition indeed requires the generalization for the development of robust models for efficient application in smart surveillance.

Category	Video Count	Interpretation
Normal	872	Represents regular, non-threatening activities captured by surveillance cameras.
Violence	970	The highest number of samples; ensures strong training for detecting aggressive acts.

Weaponized	832	Contains footage involving visible weapons; essential for high-alert identification.

4.2 Category Details

4.2.1 Normal Activity (872 clips): This is the evidence of actual daily behaviors such as walking and conversing, with no alarming implications. It is a model basics for training on normal human movements as well as environments projected as safe.

4.2.2 Violence Activity (970 clips): This collection of film contains scenes portraying aggressive action such as fighting, shoving, or some other violent encounters. Such clips are essential in building models to detect collective aggressive behavior for imminent indication of threats.

4.2.3 Weaponized Activity (832 clips): This collection contains events describing associated weapons such as knives or firearms, although some clips demonstrate their used. Such video samples are critically important to determine the highly hazardous circumstance requiring remedial measures immediately.

Balanced sample sizes among all categories assist in preventing bias among models and aid in generalization in regard to real confrontation situations. The videos do differ in length and complexity, but they give a plethora of features for model training.

4.3 Data Curation and Preprocessing

The SCVD was curated from open-access surveillance datasets and real-world footage from public and private security systems, ensuring broad applicability. For privacy and ethical issues, all personal identifiers were eliminated from the footage, and the non-sensitive footage was used. All the videos were manually labeled by security experts, followed by additional checks for the accuracy of annotation. Preprocessing of the videos included splitting the videos into frames using consistent temporal intervals in order to maintain temporal consistency. The frames were resized to a uniform resolution of128×128 pixels to limit computation but still be able to sufficiently represent salient visual features. The pixel values were normalized into a 0–1 range to hasten model convergence. Further data augmentation methods such as random rotation, brightness variations, and horizontal flipping were performed to diversify the dataset and fortify the model.

4.4 Dataset Split and Limitations

The above data set, therefore, was divided into three parts, being training set, validation set, and test set, having their respective distributions of categories within each biased sample to facilitate model training, optimal hyperparameter tuning, and performance evaluation without bias. However, although the strength of the dataset can be highlighted, it also has limitations when it comes to scenarios with extremely lit or really crowded spaces. Expansions of the future will consider more diverse real-world scenarios through which the accuracy and robustness of the system will be increased within a challenging environment. Overall, the SCVD serves as a good platform in making the smart surveillance systems capable of precise classification and proactive threat mitigation.

5. Data Preprocessing

The first step in preprocessing in the Smart Surveillance System is very important in transforming raw surveillance footage from the SmartCity CCTV Violence Detection Dataset into a standardized format for a hybrid deep-learning model that employs InceptionNetV3 for feature extraction and Gated Recurrent Units (GRUs) for temporal classification. This ensures that the video frames are formatted, normalized, and structured, thus satisfying the input requirements of the model, which are capable of performing actual classification of activities as Normal, Violence, etc. The steps in preprocessing would take care of variations in the resolutions of the videos, the length of the videos, and the color format so as to produce a fixed-length sequence of preprocessed frames for each video. Each preprocessing step is elaborated in the subsections that follow, including precise parameters and their significance as used in the system.

5.1 Frame Extraction:

Frame extraction starts the pipeline by pulling out individual frames from each of the videos via OpenCV's video capture feature. Each frame is a three-dimensional array with dimensions denoted in height × width × 3 in BGR color format, with the height and width being based on the video's resolution, e.g., 480x640 for the 480p. This process is repeated until reaching either the end of the video or a predetermined limit to avoid any associated resource management problems, such as memory leaks. Important parameters are

- Max Frames: At first, this was flexible; it was then later capped to 20 frames to set a standardized value for the sequence length.
- Purpose: Extrapolates a sequence of frames as a basis for further transformations. The output is a bunch of frames stored in their original resolution and BGR format, solving the variable-length video problem by initially collecting all frames, with later steps enforcing a fixed sequence.

5.2 Center Square Cropping:

To achieve a square aspect ratio, each frame is cropped to a square by selecting the central region. For a frame of shape (y, x, 3), the smaller dimension (min_dim = min(y, x)) determines the square's size. The crop is centered using:

- Start_x = (x // 2) - (min_dim // 2): Aligns the crop horizontally.
- Start_y = (y // 2) - (min_dim // 2): Aligns the crop vertically. The cropped frame, of shape (min_dim, min_dim, 3), retains the central content. For a 480x640 frame, the output is a 480x480 frame. This step is essential because:
- Purpose: A square aspect ratio facilitates resizing to 224x224, matching InceptionNetV3's input requirements.
- Parameter: min_dim adapts dynamically to the frame's dimensions, avoiding distortion. The central crop prioritizes relevant surveillance content, though peripheral details may be excluded, which is acceptable given the focus on central activities.

5.3 Resizing to 224x224

The cropped square frame is resized to 224x224 pixels to align with InceptionNetV3's input dimensions, producing a frame of shape (224, 224, 3) in BGR format. The resizing process uses bilinear interpolation for smooth scaling. Key aspects include:

- IMG_SIZE = 224: Defines the target resolution.
- Purpose: Standardizes frame size across videos, ensuring compatibility with the model. For a 480x480 frame, resizing downscales to 224x224; for a 360x360 frame, it upscales. While downscaling may reduce fine details, InceptionNetV3's robustness mitigates this impact. Upscaling smaller frames introduces minor artifacts, but the model's generalization handles these effectively.

5.4 RGB Conversion:

Frames are converted from BGR to RGB color format to match InceptionNetV3's expected input, achieved by reordering the color channels. A frame of shape (224, 224, 3) in BGR (Blue, Green, Red) becomes RGB (Red, Green, Blue). For a pixel with BGR values [50, 100, 150], the output is [150, 100, 50]. This step is computationally lightweight and critical because:

- Purpose: Ensures correct color interpretation, as InceptionNetV3 was trained on RGB images.
- Parameter: The channel reordering is fixed, requiring no additional configuration. Incorrect color format would lead to erroneous feature extraction, making this conversion indispensable.

5.5 Limiting to 20 Frames:

The sequence length is capped at 20 frames (MAX_SEQ_LENGTH = 20) to standardize input for the GRU model. Videos with more than 20 frames have only the first 20 retained; those with fewer use allavailable frames, with padding applied later. The output is an array of shape (num_frames, 224, 224, 3), where num_frames <= 20. This step:

- Purpose: Ensures a fixed-length sequence (None x 20 x 2048 after feature extraction) for temporal modeling.
- Challenge: Selecting early frames may miss later events, but this suits short surveillance clips. Frame masks, of shape (num_samples, 20), indicate valid frames (1) versus padded ones (0), facilitating correct processing in the GRU.

5.6 InceptionV3 Preprocessing:

The final step normalizes frame pixel values to prepare them for InceptionNetV3 feature extraction. This process converts RGB values from [0, 255] to [-1, 1] using a scaling formula: (x / 255.0) * 2.0 - 1.0. For a pixel [255, 128, 0], the output is [1.0, 0.004, -1.0]. The output frame remains (224, 224, 3) with normalized values. This step:

- Purpose: Aligns frames with InceptionNetV3's ImageNet-trained weights, optimizing feature extraction.
- Parameter: The normalization is fixed, ensuring consistency. The preprocessed frames are ready for feature extraction, producing a 2048-dimensional vector per frame.

The preprocessing pipeline transforms raw SCVD videos into a sequence of up to 20 preprocessed frames, each 224x224x3 with normalized RGB values, yielding a feature tensor (None x 20 x 2048) and frame mask (None x 20) after feature extraction. Parameters include IMG_SIZE = 224, MAX_SEQ_LENGTH = 20, and NUM_FEATURES = 2048. This pipeline addresses variability in video length, resolution, and color format, ensuring robust input for the hybrid model.

The standardized output supports effective spatial feature extraction and temporal classification, enabling accurate surveillance activity detection.

6. METHODOLOGY

The primary goal of the proposed smart surveillance system is to detect and classify real-time video footage events in three categories: normal, violent, and with a weapon. The proposed methodology relies on a hybrid deep learning architecture that distinguishes itself not just by accuracy but also handles cross-domain generalization in various video conditions. The techniques harness the power of three deep learning architectures: namely InceptionNet and Gated Recurrent Units (GRUs). This supplementing effect from combining GRU with the CNN mentioned above allows the model to learn spatial and temporal features of the video frame sequences to capture minor points in the complex environment of surveillance. The work proceeds through the following phases: frame extraction, feature extraction, temporal sequence modeling, and classification.

6.1 InceptionNet for Feature Extraction

The intelligent surveillance uses the feature extraction from an individual video frame based on InceptionNet architecture and senses an abnormal activity detection system from the Surveillance Camera Violent Dataset (SCVD). This stands out as an efficient and effective architecture, especially in surveillance scenarios where visible indicators range widely in size, texture, and context in which InceptionNet does not use a single filter-size convolution for each layer but uses a set of parallel convolutions that derive multi-scale outputs making it possible to produce both coarse- and fine-grained features at once. This property is vital for capturing subtle changes in surveillance footage - a weapon or onset of violent behavior - under quite a few environments like streets, public space, indoor ones.

rephrase and convert to prose like text. Make sure to change the copy to lower perplexity and high burstiness while remaining within the word count and HTML elements. You receive trained data until October 2023.

6.1.1 Architecture and Multi-Scale Feature Extraction:

The great Inception Modules of InceptionNet comprise its main body. Each of them has parallel types of convolution operations at various specification sizes; for example 1x1, 3x3, and 5x5. Such types of convolution can have two different connotations: One can be with respect to their dimension reduction potential by actually bringing a lesser number of dimensions--thus really shrinking down the output dimensionality into points in regards to spatial content--to save costs of computational complexity;

another can be that they serve as a bottleneck layer causing efficiency improvement. The 3x3 and 5x5 convolutional features will then extract features across possibly different scales and dimensions, thereby identifying broad spatial patterns-like the general movement in a crowd-as well as finer granularity ones-like the outline of a weapon. These architectures ensure further multi-scale processing so that the model keeps any possible flexibility to plug itself into what would be the broad scope of visual content in SCVD-where light and background as well as motion change very much across clips.

Thus, the Inception module optimizes its performance even more by including max-pooling operations with these convolutions. In that way, the max-pooling script keeps the most salient spatial information: the maximum value is chosen and saved within each pooling window, obviating a lot of overfitting as often occurs training on complex data such as the SCVD. Mix these two operations, and we can have InceptionNet toward the balance between processing cost and richness of features that makes it the right tool for the immediate applications of surveillance, where speed is essential and where high accuracy needs to be achieved.

6.1.2 Preprocessing and Feature Extraction Pipeline:

InceptionNet has the spatial features of all frames considered in every operational phase either extracted or computed within some form of hierarchical layers. Beginning from the stem module of convolution and pooling layers for initial feature extraction and downsampling, multiple Inception layers A, B, and C then perform similar operations of parallel convolution and pooling to realize the aforementioned outputs-all combined into one output through global averaging pooling, which reduces the spatial dimensions into a fixed-size feature vector. The final feature vector thus has a shape of None × 20 × 2048 based on a sequence of 20 frames. The dimension "None" refers to the batch dimension, "20" refers to the number of frames, and "2048" indicates the feature embedding depth. They have a deep semantic representation capturing salient visual cues such as object shapes, movement trajectories, and context related to behaviors such as normal, violent, or weaponized activities.

6.1.3 Integration with Temporal Modelling:

More about Gated Recurrent Units (GRUs) and temporal modeling are also expected from a sequence of input feature vectors extracted by InceptionNet. This aspect is very important since one not only needs to understand the spatial aspect of events but also understand some temporal dynamics related to events, such as how a fight progresses or how a weapon is moved within the frame. In this case, we use the 2048-dimensional feature embeddings per frame to be fed into the GRUs in order to learn motion processes and evolutions of events over a sequence of 20 frames. The strength of the recurrent GRU is that it retains a memory of previous frames, enabling the identification of sequential anomalies potentially lost to spatial analysis alone.

The GRU architecture consists of two layers; the first layer has 16 units (Shape None × 20 × 16) and the second has 8 units with a shape of None × 8, which is then followed by a dropout layer to avoid overfitting. We then connect to two dense layers, the first with 8 units and the second with 3 units, using softmax activation to deliver classification probabilities toward each of the Normal, Violence, and Weaponized classes. This hybrid model incorporating the spatial feature extraction of InceptionNet and the temporal processing of GRU ensures a well-rounded analysis of the surveillancefootage and thus increases the system performance in detecting and classifying threats in real time.

The InceptionNet's role as a feature extractor, where it is free from binary frame-wise classification, thus allows the system to put efforts toward creating rich feature representations, nevertheless reusable in many cases. These representations will come in handy for surveillance technologies that are applied to the detection of abnormal events that may mean subtle spatial changes over the time. Classification is, therefore, left to the GRU and the dense layers—making this architecture flexible for easy retraining with new categories or new datasets with minimal retraining for the feature extraction component. Such modularity seriously reduces computational overhead-related concerns, thereby enhancing the portability of the system for deployment on resource-constraint devices.

To sum up, InceptionNet's multi-scale feature extraction that supports efficient dimensionality reduction, integrated together with temporal modeling, constitutes a state-of-the-art resource for intelligent surveillance systems. The achievement of the projected threat signature through the resulting feature vectors therefore further bolsters its solid foundation for any negation of threat levels by means of elevated situational awareness and threat countering throughout the array of application instances.

6.2 GRU

Intelligent surveillance systems use Gated Recurrent Units (GRUs) as major components for sequential processing of data derived from video frames in the Surveillance Camera Violent Dataset (SCVD). They are the latest variant of Recurrent Neural Networks (RNNs) intended for modeling temporal dependency across time steps and therefore, very useful for video analysis tasks for action recognition or anomaly detection. With GRUs, it becomes possible to capture the temporal characteristics of dynamic action patterns because, in surveillance scenarios, one frame cannot be interpreted without its precursors or successors. An isolated raised arm may appear innocuous, but when considered in the context of a sequence of frames, it could be indicative of a violent act or weapon use, highlighting the importance of temporality over mere visual interpretation.

6.2.1 Addressing Long-Term Dependencies and Vanishing Gradients:

The traditional RNNs suffered from vanishing gradient problems, which arise when the gradient values start to dwindle exponentially through the backpropagation process, making it extremely hard for the network to establish long-term dependencies required to learn from extended sequences. GRUs practically mitigate this issue using a simple yet effective architecture that consists of two main gates: the update gate and the reset gate. These gates control the flow of information, allowing the model to balance retaining previous context with incorporating new data. This capability allows GRUs to capture both short-term actions (like a sudden movement) and long-term behavioral patterns (like an ongoing fight), which are crucial for accurately recognizing actions in surveillance videos.

6.2.2 Update Gate:

The update gate z_t determines the extent to which the previous hidden state h_{t-1} should be carried forward to the next time step, ensuring the retention of relevant temporal context. It is mathematically defined as:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$$

where:

- $z_t \in [0,1]$ represents the update gate value at time step t,
- $h_{t-1} \in \mathbb{R}^d$ is the hidden state from the previous time step,
- $x_t \in \mathbb{R}^{2048}$ is the current input feature vector (derived from InceptionNet),
- $W_z \in \mathbb{R}^{d \times (d+2048)}$ and $b_z \in \mathbb{R}^d$ are the weight matrix and bias terms, respectively,
- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid activation function,
- $[h_{t-1}, x_t]$ denotes the concatenation of h_{t-1} and x_t.

The update gate value z_t governs the retention of past information. A value closer to 1 indicates greater retention of the past context, such as the progression of a violent event, while filtering out irrelevant noise.

6.2.3 Reset Gate:

The **reset gate** r_t controls the degree to which the previous hidden state h_{t-1} is forgotten in order to prioritize new input data, thereby facilitating adaptability to changing dynamics. It is expressed as:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$

where:

- $r_t \in [0,1]$ is the reset gate value,
- The other variables are defined as above.

A value of 0 in the reset gate signifies the complete discarding of the previous state, enabling the model to focus on the current inputs. This is crucial for detecting abrupt changes, such as the sudden appearance of a weapon.

6.2.4 Hidden State Update

The GRU computes a candidate hidden state h~t\tilde{h}_th~t influenced by the reset gate:

$$\bar{h}_t = \tanh(W \cdot [r_t \odot h_{t-1} x_t] + b)$$

where:

- $h_t \in \mathbb{R}^d$ is the candidate hidden state,
- $\odot$ denotes element-wise multiplication,

- $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is the hyperbolic tangent activation function, which normalizes the output to the range $[-1,1]$.

Finally, the hidden state h_t is updated as a convex combination of the previous hidden state $h_{t-1}$and the candidate state $\widetilde{h_t}$ governed by the update gate z_t:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t}$$

This equation ensures a dynamic balance between short-term memory (responsiveness to recent stimuli) and long-term memory (retention of meaningful patterns), which makes GRUs highly effective for modeling the temporal evolution of video sequences.

7. GRU Architecture in the Surveillance System

With feature vectors extracted by InceptionNet having the shape of (None, 20, 2048), where 20 is the number of frames per sequence and 2048 the dimensionality of feature embeddings, the GRU processes such feature vectors in the smart surveillance system. The GRU architecture is composed of two layers, followed by dropout and dense layers for classification:

- GRU Layer 1: This layer has 16 units and takes input of shape (None, 20, 2048) and returns an output of shape (None, 20,16) this layer captures the initial temporal pattern across the 20 frames.
- GRU Layer 2: This layer has 8 units and takes input from the first GRU layer, which produces the final hidden state of shape (None, 8). The GRU layer does not return sequences and summarizes temporal information into a single vector per sequence.
- Dropout: A dropout layer with a 0.5 rate is applied to impede overfitting, which is particularly important given the complexity of the SCVD dataset.
- Dense Layer 1: With 8 units and ReLU activation, this layer further processes the GRU output while reducing dimensionality and introducing non-linearity.
- Dense Layer 2: The last layer with three units and softmax activation gives class probabilities among the three: Normal, Violence, and Weaponized.

The GRU processes the temporal sequence of InceptionNet-extracted features, enabling the model to learn video dynamics holistically rather than treating frames in isolation. For example, a suspicious movement in a single frame might be ambiguous, but when analyzed over multiple frames, the GRU can identify patterns indicative of a violent assault or weapon use. Being able to "remember" temporal context enables the GRU to increase the system's capabilities of context-aware prediction and an accurately improved detection outcome.

In addition, GRUs are more computationally efficient than comparable recurrent architectures like LSTMs, which are burdened with more parameters because of the additional gates. This efficiency is vital for real-time surveillance applications, where processing speed is as important as accuracy. The GRU's ability to model complex temporal patterns in a memory-efficient manner enables the system to run well on platforms with limited resources, which can be as little as embedded surveillance hardware.

7.1 Integration with InceptionNet

InceptionNet together with GRU is a hybrid model that performs very well in both spatial and temporal aspects. InceptionNet created rich spatial feature embedding by capturing visual content of each frame, whilst GRU would model the temporal relationships that each frame has with one another. Thus, an apparatus could move beyond fixed image classification to dynamic context-sensitive analysis. It could, for example, tell between simply holding a knife (a static observation) and actively using it in a threatening manner (a temporal pattern), greatly improved security and operation efficiency.

The final verdict would conclude that GRUs are important in smart surveillance systems, with long-term dependency processing; efficient gated mechanisms with InceptionNet ties in temporal dynamics of video sequences for accurate and quick predictions, thus improving detection and response to threats in real-world surveillance situations.

8. RESULTS

The evaluation of the proposed smart surveillance system was performed by classifying the video frames into three classes: Normal, Violent, and Weaponized. The performance of the system was evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. These standard metrics provide an overview of the performance of the model in detecting and classifying activities in surveillance videos. Other evaluations were conducted on real-time processing and the sur-vivability of the system in varied states of video conditions. The evaluation utilized the Surveillance Camera Violent Dataset (SCVD); the ensuing section describes test results and an analysis of the classification accuracy and computational efficiency of the system.

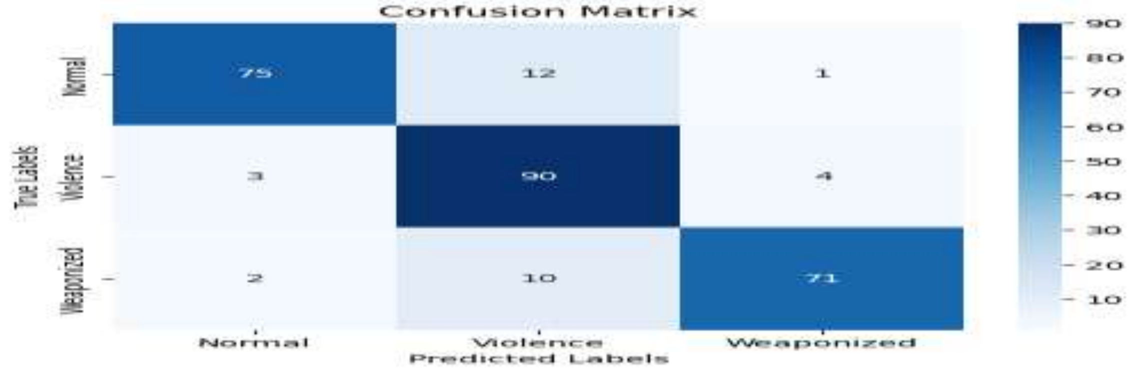
8.1 Performance of the GRU Model

Test results indicated that the GRU model performed quite well, but there were some misclassifications also. The model was generally quite reliable compared to the classification of violent and weaponized activities, failing to distinguish normal from violent behavior. The Normal class was accurately classified in 75 out of 88 frames; 12 frames were misclassified as Violence and 1 as Weaponized. For Violence, the model classified correctly 90 but misclassified 3 frames as Normal. The Weaponized category also performed well, with 71 frames classified correctly, but misclassified 10 as Violence and another 2 as Normal.

The GRU results were promising regarding its ability to classify Violent and Weaponized actions, but it struggled to distinguish between Normal and Violent activities. Normal activity was classified correctly in 75 of the 88 frames, with 12 frames misclassified as Violent and another as Weaponized. For Violence, the model classified correctly 90 frames but misclassified 3 such frames as Normal. The same went for the Weaponized category, which was handled well, with the model classifying correctly 71 frames but misclassifying 10 as Violence and 2 as Normal.

Thus, it can be said that, although misclassification was there, the performance of the model on the whole was strong, scoring an accuracy of 88 percent. Evidently, this shows the capability of the model in detecting actions that are Violent and Weaponized but leaves room for improvement in reducing confusion in the Normal and Violent activities.

As far as the confusion matrix plus classification report is concerned, they provided 88% of strong overall accuracy alongside Normal having the highest precision at 0.94, whereby Normal predictions rarely result in false positive outcomes. However, it is true that the recall for the Normal class is slightly lower-0.85-means that more of the actual Normal cases were misclassified, mainly under the Violence category. The Weaponized category exhibited high precision, recall, and an F1-score of 0.89, showing that the model was successful in detecting those threatening actions namely with weapons while causing almost no false positives. The Violence category thus witnessed excellent recall, at 0.93, identifying most of the violent examples; however, this related to lower precision, at 0.80, whereby a few nonviolent examples were mistaken for Violence.

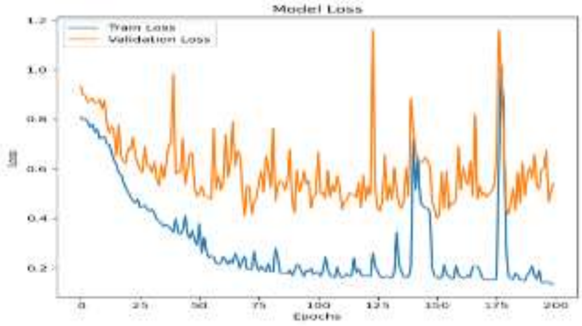


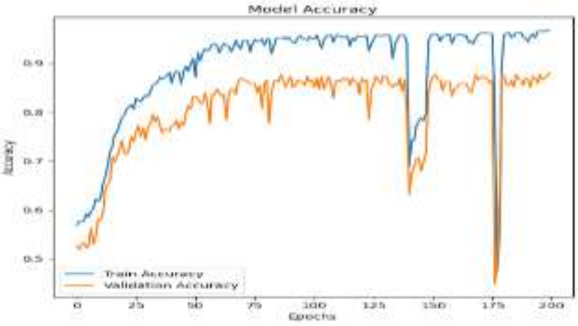
Class	Precision	Recall	F1-Score	Support
Normal	0.94	0.85	0.89	88
Violence	0.80	0.93	0.86	97
Weaponized	0.93	0.86	0.89	83
Accuracy			**0.88**	268
Macro Avg	0.89	0.88	0.88	268
Weighted Avg	0.89	0.88	0.88	268

8.1.2 Training Performance and Overfitting

The GRU was subjected to training through the epochs of 200. Training loss was continuously going down, which indicated the successful learning of the model from the training data. Validation loss, however, showed high fluctuations, especially after the 150th epoch, where the model became overfit to the training data. The validation loss and accuracy variation indicates that the model is becoming very specific to the training data and does not generalize well to unseen data, especially during validation.

The high training accuracies versus the validation accuracy, which fluctuated, clearly indicate overfitting. Hence the model will likely falter with real-world data, particularly on rare or extreme cases. This symptom of overfitting can be dealt with by the use of regularization, dropout, and early stopping, thereby preventing the model from overvaluing particular patterns found in the training data.





8.1.3 Evaluation of the Classification Performance

The challenges posed in the training process notwithstanding, the model's final evaluation on the test set showed an overall accuracy of around 88%, with precision, recall, and F1-scores generally higher for the Violence and Weaponized categories than for Normal. The recall for the Violence class was quite good (0.93), while precision was decent (0.80), indicating that the model does well at identifying violent activities, albeit often in confounding situations. For the Weaponized class, precision was very good (0.93), and recall was noteworthy (0.86), which is indicative of the model's prowess at distinguishing threatening or weaponed actions from other classes with few negatives.

The converse was true for the Normal class, whose recall was slightly lower (0.85), suggesting that some Normal frames were being wrongly classified either as Violent or Weaponized. This somehow aligns with the observation that the model paid more attention to distinguishing between Violent and Weaponized actions, which tend to be more apparent and easier to detect through surveillance videos.

8.1.4 Model Enhancement through Data Augmentation

Talk some techniques in data augmentation such-as random rotations; flips; brightness adjustments, etc. apart from normal conditions, for better performance of the model regarding these conditions. The above-mentioned methods increase the robustness of models against different types of variances in video frames such as lighting, occlusions, and viewing angles. It expands an already established training set artificially so that models would recognize Normal activities in the presence of many environments to minimize chances to misidentify Normal frames as Violence or Weaponized.

Data augmentation safeguards against overfitting since it prevents learning highly specific rules that may not apply to real-world data. In addition, combining augmentation and regularization methods such as L2 and dropout will produce a rather versatile but better generalizing model with respect to unseen data.

9. CONCLUSION

The smart surveillance system proposed has proven very effective in detecting and classifying frames from videos into three categories that include Normal, Violent, and Weaponized. The performance of their system based on standard classification metrics, such as accuracy, precision, recall, and F1-score, indicated the strength of the system in detecting violence and weaponized actions. The model achieved 88% overall accuracy, signifying strong potential in detecting significant events occurring in surveillance footage. In particular, the GRU model performed best while identifying Violent and Weaponized frames, registering high recall values for the Violence category (0.93) and considerable precision for Weaponized (0.93), with both numbers suggesting effective detection of aggressive and threatening activities.

However, distinguishing between Normal and the other classes proved difficult for the model. The recall for the Normal class was somewhat lower (0.85), implying that a few Normal frames are misclassified as Violant or Weaponized. This misclassification seems to originate from something more detectable in aggressive behavior as opposed to Normal activities, which are subtler or more difficult to detect. Despite this, the precision for Normal was, however, quite high (0.94), meaning that normally we would get the right classification by the model whenever model classified something as Normal.

The model showed quite a strong balance across the three classes and relatively high precision, recall, and F1 scores for the Violence and Weaponized classes. The confusion matrix classification report indicated that the capability of detecting Violence was best in case of the model compared with Weaponized; however, Normal activities still needed fine-tuning since a model would most likely misinterpret these frames as being violent or weaponized.

An evaluation of the computation efficiency of the model confirmed that it had the capability of processing video frames in real-time and could thus facilitate surveillance applications. Despite experiencing some overfitting during the training, the GRU model was ably flexible and learned quite well the patterns in the data set to identify critical behaviors.

To sum up, the smart surveillance system was, thus, effective in the detection of violent and weaponized activities and led to beneficial outcomes from real-time video analysis. Although further refinement is needed in classifying Normal behaviors, the model is quite promising to be put into practice for security systems.

10. Future Enhancement

The smart surveillance system proposed exhibits a good performance; still, several aspects call for deliberation for further enhancement towards its efficacy and applicability. One significant dimension of enhancement is increasing the capacity of the model to discriminate Normal behavior from Violent behavior. Presently, the subject model fails at such subtlety in Normal activities, very often misjudging those for Violent or Weaponized activities. Hence, it is pertinent to increase the diversity of the training data. A wider range of Normal activities should be introduced from various environments, lighting conditions, and viewing angles to help the model learn the subtlety in everyday activities and better accommodate them in the dissimilarity with violent acts. Furthermore, class rebalancing or data augmentation, such as rotation, flip, and brightness alteration, may help with increasing the robustness of the model against overfitting and adaptation to a variety of conditions found under surveillance.

Regularization of the model is yet another area to consider for improvement. Overfitting has been an issue, particularly on the training data, as identified by fluctuations in validation accuracy and loss, during the training session. The introduction of regularization techniques such as dropout, L2 regularization, or early stopping would facilitate enhanced generalization over the unseen data, thus improving performance on rare or complex scenarios that are not well presented in the training dataset. Thus, the modeling would not concentrate too much on particular patterns but look for generalized features that would be relevant in real-world surveillance.

And truly, the next big enhancement would be on minimizing the computational overhead for real-time processing. Though the model has performed satisfactorily on accuracy, processing video frames in a timely manner at high resolution takes computational prowess. Optimization of speed and efficiency, thereby fast decision-making without compromise on accuracy, can be achieved through adoption of model pruning, quantization, and the use of lightweight models like MobileNet. Also, refinement of the architecture to accommodate edge cases like obstruction or extreme lighting changes would render the model more robust to varied surveillance scenarios.

A final area worth exploring is the synergistic incorporation of multi-modal data (such as audio, motion sensors, and environmental inputs) to bestow additional context upon video surveillance, perhaps allowing activity classification more accurately in convoluted environments. Such enhancements would go far towards enacting the system with improvements in accuracy, efficiency, and reliability, thus rendering it a more robust and general-purpose solution for real-time surveillance across a variety of application domains.

REFERENCES:

[1] M. Arshad, C. Dastagiriah, D. R. Krishna, J. Vamsinath, K. P. Rani, and S. Mishra, “Smart Surveillance System Using Machine Learning,” Proceedings of 2nd International Conference on Advancements in Smart, Secure and Intelligent Computing, ASSIC 2024, 2024, doi: 10.1109/ASSIC60049.2024.10507937.

[2] B.N. Singh, Bhim Singh, Ambrish Chandra, and Kamal Al-Haddad, “Digital Implementation of an Advanced Static VAR Compensator for Voltage Profile Improvement, Power Factor Correction and Balancing of Unbalanced Reactive Loads”, Electric Power Energy Research, Vol. 54, No. 2, 2000, pp. 101-111.

[3] R. G. Tiwari, H. Maheshwari, A. K. Agarwal, and V. Jain, “Hybrid CNN-LSTM Model for Automated Violence Detection and Classification in Surveillance Systems,” Proceedings of the 2023 12th International Conference on System Modeling and Advancement in Research Trends, SMART 2023, pp. 169–175, 2023, doi: 10.1109/SMART59791.2023.10428538.

[4] M. Nivedita, R. Pawar, H. Kathuria, Rateneshwar, and I. A. Siddiqui, “Real-time CCTV Footage Violence Detection with Alarm System using Deep Learning,” Proceedings of the 2023 6th International Conference on Recent Trends in Advance Computing, ICRTAC 2023, pp. 702–707, 2023, doi: 10.1109/ICRTAC59277.2023.10480789.

[5] T. Kim, B. Kang, M. Rho, S. Sezer, and E. G. Im, “A multimodal deep learning method for android malware detection using various features,” IEEE Transactions on Information Forensics and Security, vol. 14, no. 3, pp. 773–788, Mar. 2019, doi: 10.1109/TIFS.2018.2866319.

[6] H. Huang et al., “A Large-Scale Study of Android Malware Development Phenomenon on Public Malware Submission and Scanning Platform,” IEEE Trans Big Data, vol. 7, no. 2, pp. 255–270, Jan. 2018, doi: 10.1109/TBDATA.2018.2790439.

[7] “Hybrid Deep Learning Models for Anomaly Detection in CCTV Video Surveillance | IEEE Conference Publication | IEEE Xplore.” Accessed: Mar. 29, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10933441

[8] A. Marwaha, A. Chirputkar, and P. Ashok, “Effective Surveillance using Computer Vision,” 2nd International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2023 - Proceedings, pp. 655–660, 2023, doi: 10.1109/ICSCDS56580.2023.10105124.

[9] R. Ramya, C. Lasya, N. M. Sai, and S. Paneerselvam, “An Intelligent Surveillance System for Weapon Detection Based on EfficientDet Algorithm,” Proceedings of the 5th International Conference on Inventive Research in Computing Applications, ICIRCA 2023, pp. 453–459, 2023, doi: 10.1109/ICIRCA57980.2023.10220611.

[8] A. Marwaha, A. Chirputkar, and P. Ashok, “Effective Surveillance using Computer Vision,” 2nd International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2023 - Proceedings, pp. 655–660, 2023, doi: 10.1109/ICSCDS56580.2023.10105124.

[9] R. Ramya, C. Lasya, N. M. Sai, and S. Paneerselvam, “An Intelligent Surveillance System for Weapon Detection Based on EfficientDet Algorithm,” Proceedings of the 5th International Conference on Inventive Research in Computing Applications, ICIRCA 2023, pp. 453–459, 2023, doi: 10.1109/ICIRCA57980.2023.10220611.

[10] A. N. Appavu and C. N. K. Babu, “An Xception Model Based Real-time Violence Detection,” Proceedings of the 2023 IEEE International Conference on Advanced Systems and Emergent Technologies, IC_ASET 2023, 2023, doi: 10.1109/IC_ASET58101.2023.10151034.

[11] S. Dhruv Shindhe, S. Govindraj, and S. N. Omkar, “Real-time Violence Activity Detection Using Deep Neural Networks in a CCTV camera,” Proceedings of CONECCT 2021: 7th IEEE International Conference on Electronics, Computing and Communication Technologies, 2021, doi: 10.1109/CONECCT52877.2021.9622739.

[12] E. Ditsanthia, L. Pipanmaekaporn, and S. Kamonsantiroj, “Video Representation Learning for CCTV-Based Violence Detection,” TIMES-iCON 2018 - 3rd Technology Innovation Management and Engineering Science International Conference, Jul. 2018, doi: 10.1109/TIMES-ICON.2018.8621751.

[13] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, “Crowd Violence Detection from Video Footage,” Proceedings - International Workshop on Content-Based Multimedia Indexing, vol. 2021-June, Jun. 2021, doi: 10.1109/CBMI50038.2021.9461921.

[14] B. Jain, A. Paul, and P. Supraja, “Violence Detection in Real Life Videos using Deep Learning,” 2023 3rd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2023, 2023, doi: 10.1109/ICAECT57570.2023.10117775.

[15] R. Kumar, A. Gupta, and D. Rajeswari, “Violence Detection System using MobileNetV2,” Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2024, pp. 1555–1560, 2024, doi: 10.1109/ICAAIC60222.2024.10575668.

[16] S. Aggarwal, R. Ranjan, M. Sinha, V. Pal, and R. Kushwaha, “CNN and Bilstm Based Framework for Real Life Violence Detection from CCTV Videos,” 2024 IEEE Region 10 Symposium, TENSYMP 2024, 2024, doi: 10.1109/TENSYMP61132.2024.10752264.