IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 March 2025, accepted 17 April 2025, date of publication 25 April 2025, date of current version 5 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3564435

RESEARCH ARTICLE

Advanced Detection of Violence From Video: Performance Evaluation of Transformer and State of the Art of Convolution of Neural Network Transformer

ABDULRAHMAN ALSHALAWI, WADOOD ABDUL, (Member, IEEE), AND GHULAM MUHAMMAD, (Senior Member, IEEE)

Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Abdulrahman Alshalawi (439106358@student.ksu.edu.sa)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project under Grant RSPD2025R1051.

ABSTRACT Safety is paramount in every aspect of human concern. In situations where violence occurs beyond the control of observers, it is essential for trained professionals to intervene. These scenarios can be detected by advanced surveillance systems equipped with cameras or sensors. The performance of these systems is highly dependent on the models they incorporate. Although various machine learning and deep learning models have been explored to address these challenges, significant work remains in handling a diverse range of images. This study aims to bridge this gap by training four models—VGG16, MobileNetV2, Yolov6 and Ensemble Model, and a vision Transformer model ('google/vit-base-patch16-224-in21k') —with tuned parameters on publicly available video frames. Our analysis established that the Transformer model achieved the highest accuracy, making it the most suitable for applications involving the detection of violent incidents. This research introduces a novel application of AI for semi-automated violence recognition, demonstrating AI's significant potential in improving public safety. It also emphasizes the superior performance of the Transformer model in accurately detecting violent frames. Semi-automated violence recognition refers to a system where AI models assist in identifying violent incidents within video footage, but human intervention is still required for certain tasks, such as verification or decision-making. The AI system automatically detects potential violent events, flagging them for further review by human operators, who then confirm or act upon these alerts. The Transformer model achieved outstanding results: for the Violence Dataset, it attained a precision of 99%, recall of 99%, F1 score of 99%, and accuracy of 99%. To further verify the performance of the Transformer model, we tested it on the Road-Anomaly Dataset, which presents a challenging scenario with imbalanced classes. For this dataset, the Transformer model achieved a precision of 98%, recall of 97%, F1 score of 98%, and accuracy of 98%. These results affirm the model's efficacy, positioning it as the top choice for real-time surveillance and public safety applications.

INDEX TERMS Deep learning, VGG16, MobileNetV2, ensemble model and transformer.

I. INTRODUCTION

To continue, it is imperative to understand a difference between opinion radicalization and action radicalization [1]. Some extreme beliefs do not always result in serious behaviors or any behaviors at all, thus calling for scientific rationalism. Also, it is essential to differentiate between nonviolent and violent actions as it is the latter which is most significant when speaking about public security purposes [2]. Thus, it is crucial to gain a deeper insight into the relationship between violent and nonviolent radicals. The best and effective way of semiautomatic identification of violent

The associate editor coordinating the review of this manuscript and approving it for publication was Olarik Surinta.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

scenes is through images or frames taken from the video adopting the modern state of art tools and technique of artificial intelligence (AI). Through analyzing visual information, AI applications can identify signs of violence, for instance, movement, or a weapon, within a short span of time. This way not only increases the speed and accuracy in the detection of violent scenes, but also allows to react instantly to possible threats, thus increasing security in public spaces. Using AI for this purpose makes it possible to maintain constant surveillance and intervene proactively should there be any violent incidences. Most current methods of violence detection involve using manual feature descriptors to differentiate between fight sequences and normal sequences which is also common in Identifying human actions Many techniques since the introduction of specific datasets for violent/fight detection have relied on the construction of artificial features representations including STIP, MoSIFT, Motion characteristics, and motion blobs used for AV analysis with blood and flame [3] Nevertheless, a limited number of studies has used deep learning methods like 2D-CNN, 3D-CNN and C3D [4]. Furthermore, deep representation models that use transfer learning techniques. To address the violent/fight detection challenge in the violent action recognition domain [5]. The principal objective of the proposed work is to accurately and immediately detect violent situations and promptly alert security departments with the extracted location data using smart CCTV cameras. This system is designed to enhance public safety by providing real-time notifications of potential threats, allowing for swift intervention and response. Besides, by incorporating this technology in smartphones, children cannot access or watch any violent materials they are prohibited from. Given the fact that the use of complex detection algorithms is integrated into the system, it will be possible to filter out the occurrences of violence, so that young viewers will not be exposed to the negative influence of the advertised material. It not only helps strengthen security measures in the areas that are open to the public but positively affects more controlled and safer virtual space for children. To address the current limitations, a four deep learning models are developed VGG16 model with tuned hyperparameters, MobileNetV2 model with tuned hyper parameters, Ensemble-Model integrating MobileNetV2 and NASNet (Neural Architecture Search Network) and Transformer model with tuned hyper parameters to detect violence and not-violence images from video frames. Proposed work implemented all above models and provide comprehensive comparison and found Transformer is selected as best model while Ensemble-Model also achieved best results. Selected model is not only best with respect to rest of model but also achieved best results from similar studies. The complementary strengths of these models are expected to enhance performance in this classification task. To confirm the effectiveness of the proposed method, experiments were conducted out using publicly accessible datasets. To validate the effectiveness of the proposed method, experiments were conducted on publicly available dataset, demonstrating that the model consistently delivers stable performance. The main contributions of this paper are given as:

- The proposed work employs artificial intelligence (AI) for the semi-automated recognition of violent scenes using image or video frames, highlighting AI's efficiency in identifying violence and its potential for preventive measures in public safety. It describes the use of deep learning models, including VGG16, MobileNetV2, an Ensemble Model with MobileNetV2 and NASNet, Yolov6 and Transformer models, all developed based on tuned parameters, to identify violent or non-violent images from video frames.
- The models 'suggested responses to the violent scenario aim to improve public safety. Security cameras from this model can identify a situation like this and alert security management when it occurs. In the interim, authorities can be notified about the situation and use automatic loudspeaker announcements and flashlight activation.
- Additionally, it advocates for the use of violence-detecting technologies in smartphone screens to prevent young audiences from watching violent content, ensuring a healthier atmosphere for young viewers.

This study is a significant contribution to the literature, offering a robust comparison of the proposed deep learning models and establishing their reliability through experiments using public-domain datasets. Both the Transformer and Ensemble Models demonstrated superior performance, providing the best scores in violence detection.

The rest of the paper is structured as follows: Section II outlines the current anomaly classification models and discusses their drawbacks. In Section III, we introduce the proposed model, while Section VI discusses the results. Lastly, Section V summarizes the proposed approach and outlines future directions.

II. RELATED WORK

Due to the increasing occurrence of anomaly in different domains, the perception of human behavior particularly, detection of violence has become an influential viewpoint in investigations concerning Computer Vision (CV). However, owing to interaction with other people and other difficulties to identify situations that are connected to definite incidents and possess certain characteristics, the identification of violence is one of the most complicated tasks in CV [6].

For detection of violent scenes, most of the work has employed the technique of machine learning [7], [8] and deep learning [9], [10], [11] These advanced techniques make it possible to detect violence through different types of data factors, and this makes the detection system more accurate and efficient. Particularly, much effort has been made to focus on text mining for identifying the measure of aggressive attitudes or violent language. In this case, textual content is analyzed and comprehended to establish relationships and indicators of violence. Current research has provided complex models that can crawl through thousands of posts from social media,

forums, and other communication channels and pick on any signs of aggressive communication [12], [13], [14].Some authors have centered on real-time violence detection concerns. and achieved notable results. In [15] Surveillance video cameras were employed to collect unique data in real-time. By analyzing changes in the magnitudes of flow vectors, they employed a descriptor called violent flow to gather statistics from short frames in their dataset. This technique resulted in an accuracy of 82.9%.

In further experiments utilizing ensemble models, the authors in [16] created ensemble learning models to identify a cyberbullying dataset sorted into two categories: 'offensive' and 'non-offensive' tweets, consisting of approximately 9093 tweets. They achieved a the highest accuracy of 96% in their experiments. They developed a single-level ensemble by integrating seven machine learning models. Furthermore, they developed a double-level ensemble that combined three machine learning models into a new model, four models into a new model, and then combined these two innovative models into a final ensemble. They found that the double-level ensemble performed the best. While several studies have used ensemble models to identify cyberbullying, other studies have used more conventional machine learning methods. This shows that ensemble learning improves performance and manages large-scale datasets more efficiently than typical machine learning models.

In a comparable research paper [17], bullying comment and post training and testing were conducted using two machine learning algorithms, namely the support vector machine and the Naïve Bayes algorithms. The following outcomes were evidenced that these classifiers achieved an accuracy of 71 percent in recognizing bullying. 25% and 52. 70%, respectively. In the past few decades, Transformers have emerged to become very popular achieving excellent performances in almost every domain. Because of this catalytic rise in Transformer-related creations, researchers have been motivated to fine-tune these architectures for various purposes activities. For instance, [18] proposed the Vision Transformer (ViT) for the image recognition task which splits the binary image input to patches and feeds it to the Transformer as sequence of tokens. This method has established benchmarks on several standard and widely used datasets. Extending on the ViT model, [19] proposed the Video Vision Transformer (ViViT) wherein the authors split the videos into spatial and temporal patches and feed them through a ViT framework. ViViT has a highly defined temporal attention mechanism that identifies the interaction between two frames and is implemented to be the most advanced in video recognition that shows great results in Kinetics-400 and Something-Something V1 and V2.

In another related advancement, [20] created the Swin Transformer, initially designed for image recognition but quickly becoming a benchmark model that surpasses other contemporary architectures. Further extending its capabilities, they developed the Video Swin Transformer, which processes spatiotemporal video patches as token sequences. This adaptation has achieved similar high performance across various video recognition tasks, confirming its effectiveness. Work [21] applies machine learning techniques to skeletal data from depth sensors to identify child physical abuse. However, such sensors are not commonly found in most schools or public places, where surveillance cameras are more widely available. In practice, surveillance cameras are often used mainly as evidence rather than for preventing incidents, as continuous 24-hour monitoring requires significant resources. Video surveillance cameras can be used to discover abuse of children early. Early violence detection using deep learning has become more popular. significantly and has shown good performance. However, deep learning models come with certain limitations, such as requiring significant computational power and vast amounts of domain-specific data. Creating large, labeled datasets is a labor-intensive and time-consuming process, posing a significant obstacle when creating deep learning models from the ground up for a given domain. Transfer learning becomes a useful strategy to overcome this issue. A source network that has been pre-trained on a sizable dataset is re-trained on a target domain-specific dataset in transfer learning. This approach removes the requirement for to produce large datasets and train models from scratch.

III. PROPOSED METHODOLOGY

The proposed work consists of the development and evaluation of four models: VGG16, MobileNetV2, YOLOv6, an Ensemble-Model, and a Transformer model were used, having fine-tuned each model based on the hyperparameters as illustrated in Figure 1 These models were trained using a set of images showing violent scenarios and actions. Based on the outcomes of these models, an additional analysis was carried out to ascertain which of the models provided a better means of identifying violent scenarios. The process of evaluation is to find the model that yields the best performance by testing its ability to predict different categories of violent imagery for practical use.

The VGG16 neural network architecture, shown in Figure 2 is named for its 16 layers with tunable weights. Specifically, VGG16 contains 13 convolutional layers, 5 max pooling layers, and 3 dense layers. Despite the architecture consists of 21 levels, only 16 of them have learnable weights. VGG16 takes input tensors of size 224×224 or 244×244 with three RGB channels. VGG16 stands out for its consistent usage of convolutional layers with 3×3 filters and a stride of one, as well as max pooling layers with 2×2 filters and a stride of two. This design choice reduces the number of hyperparameters and helps preserve spatial information. Another notable characteristic is the uniform arrangement of its convolutional and max pooling layers. The first convolutional layer, Conv-1, has 64 filters, Conv-2 has 128 filters, Conv-3 has 256 filters, and Conv-4 and Conv-5 each have 512 filters [22], [23] After the convolutional and pooling layers, the output tensor is flattened to a one-dimensional vector. This flattened tensor is then input into two dense hidden layers,

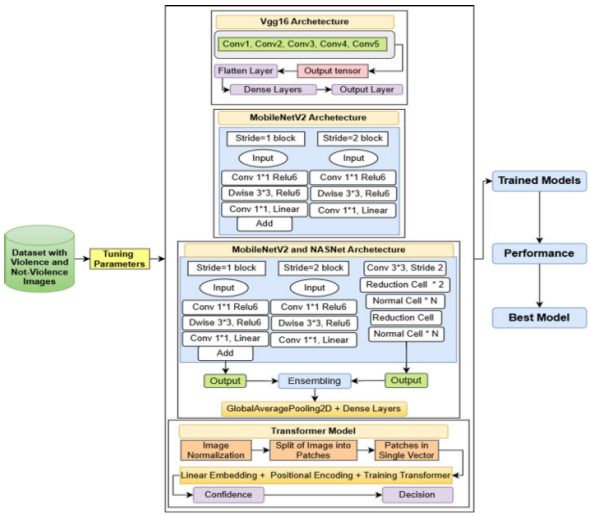

FIGURE 1. Process for selection of the best ModelVGG-16.

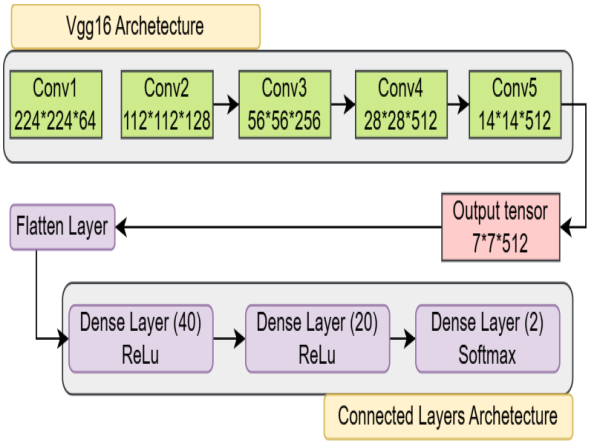

FIGURE 2. Proposed work based on VGG16.

each with 40 and 20 neurons, which use the rectified linear unit (ReLU) activation function. The output layer is made up of two neurons with a softmax activation function, which provide two output vectors: [1, 0] for the ''Not-Violence'' category and [0,1] for the ''Violence'' category.

A. MobileNetV2

MobileNetV2 offers two separate sorts of blocks: a residual block with stride 1 and a shrinking block with stride 2. Each block has three layers. The initial layer in both types is a 1×1 convolution with ReLU6 activation. The second layer employs depthwise convolution, and the final layer is another 1×1 convolution without any non-linearity. It is asserted that applying ReLU again would restrict deep networks to the capabilities of a linear classifier on the non-zero volume segment of the output domain. An expansion factor, denoted as t, is incorporated with a fixed value of t=6 in the primary experiments. For example, if the input has 64 channels, the internal output will have $64 \times t = 64 \times 6 = 384$ channels. The first block's inputs consist of three-channel Images are

TABLE 1. Architecture for MobileNetV2.

Layer	Output	Stride	Filter	Config
Conv2D	224x224 x 32	-	32	Relu-KS: 3x3
Max-Pooling	112 x 112 x 32	2	-	KS: 2x2
Conv2D	113 x 113 x 96	-	96	Relu-KS: 3x3
Max-Pooling	56 x 56 x 96	2	-	KS: 2x2
Conv2D	57 x 57x144	-	144	Relu-KS: 3x3
Max-Pooling	28x28x144	2	-	KS: 2x2
Conv2D	29x29x192	-	192	Relu-KS: 3x3
Max-Pooling	14x14x192	2	-	KS: 2x2
Conv2D	15x15x576	-	576	Relu-KS: 3x3
Max-Pooling	7x7x576	2	-	KS: 2x2
Output Module	62720	-	-	-

224×224 in size, with 3×3 kernel sizes and 32 filters, utilizing ZeroPadding2D padding. The spatial size of the data representation (width and height) is then lowered using a Max Pooling (MP) layer with a step size of two. This reduction largely reduces image size since more pixels correspond to more parameters, resulting in a big amount of data. The subsequent blocks process the data similarly, but the filters used are 96, 144, 192, and 576 as shown in Table 1 A feature map is created by combining all block outputs and is then fed into the output module.

B. YOLOv6

YOLOv6 is a modern object detection system specifically designed to deal with real-world problems in industrial settings. The architecture starts with an EfficientRep Backbone that utilizes re-parameterizable RepBlocks during training. These RepBlocks are then converted into simpler RepConv layers to enhance training speed. These features flow into the Rep-PAN neck, which specifically designed to merge multi-scale information using up sampling (''U'') and channel concatenation (''C''), making the model to locate both small and large objects. These features are refined at every stage through repeative RepBlocks and convolutional layers, ensuring the network preserve important details. Later on, three Efficient Decoupled Heads are incorporated, each working with a different detection scale and splitting its tasks into classification (''cls.'') and regression (''reg.'') to avoid confusion between recognizing and locating the objects. TO cut down complexity and speed up training, YOLOv6 directly predicts the distance of a bounding box from a reference point. During training, all multi-branch structures are collapsed into efficient 3×3 convolutions using re-parameterization, making it great for real-time applications. Refined loss functions like VariFocal Loss and SIoU are used

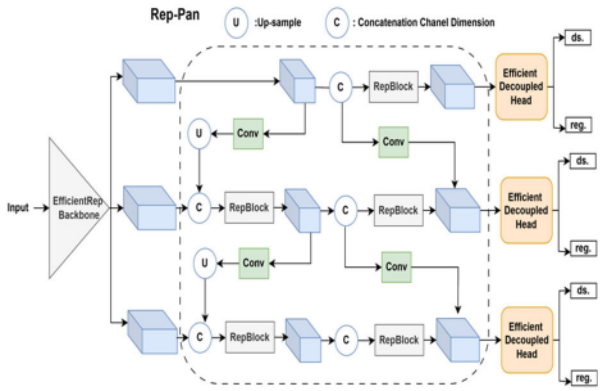

FIGURE 3. The YOLOv6 framework.

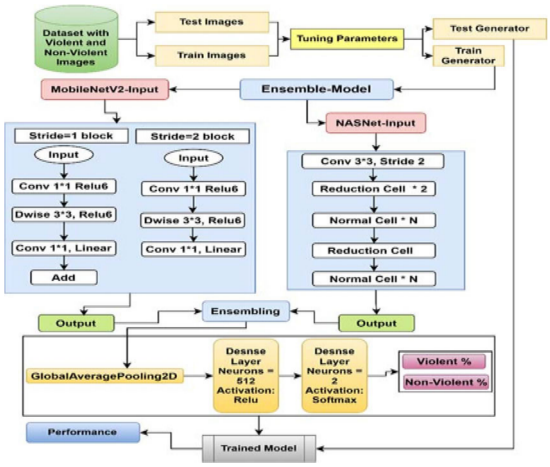

FIGURE 4. Ensemble-model workflow.

with Task Alignment Learning to keep the predictions more accurate [24], [25]. The whole architecture of YOLOv6 is depicted in Figure 3.

To convert Yolov6 object detection output into frame-level violence or non-violence decisions, we depended on the identity of specific object classes from Coco Dataset. Depending on the domain knowledge, some objects - such as knives, guns, or other potentially dangerous objects - were labeled as an indicator of violent activity. When Yolov6 detected in one of these dangerous items, the frame was classified as the violence category. Conversely, if it is discovered that no objects match the predetermined people as dangerous, the frame was classified as non-violent. This method provided a direct post-processing approach yet, using the presence or absence of dangerous objects as a base to classify each frame.

C. SEMBLE-MODEL CONSISTED OF MobileNETV2 AND NASNET

The proposed methodology comprises three steps. In the first step, the ensemble model receives an image vector from the input layer. Then, it performs convolutions and max pooling, and the resultant vector is passed to global average pooling, as illustrated in Figure 4.

In the proposed artists' contingent model, the output of Mobilentv2 and Nasnetlarge was added using a feature-condensation approach. In particular, both models were pre-trained on the imagenet and used without their top classification layers. After their output, the same input image was applied to the tensor, both were concurrent using the containet () layer to merge the deep features extracted from the architecture. This joint feature representation was then passed through a series of additional layers, including a globalaveragepooling2D layer and a fully associated (dense) layer with a globalavergepooling2D layer and relu activation, before producing the final prediction through a dense layer with softmax activation. This indicates that instead of average or weighing the output, the model used a dedicated classifier trained on fused features to make the final classification. Although the model is trained on a custom dataset, using imagenet pre-trained weight is a common and effective transfer learning strategy. This allows the model to take advantage of the feature representation already learned, which can significantly promote performance and reduce training time, especially when the target is limited in dataset size. This attire strategy provides many advantages: it combines the complementary strength of both models - the computational efficiency of Mobilentv2 and the deep representative power of the nasnetlarge - a rich and more diverse feature reconsidered into the set. As a result, it enhances the ability to reduce the risk of overfitting compared to better normalizing the model's ability, improving accuracy and using single architecture alone.

D. USED MODELS

The ensemble uses two models: MobileNetV2, as described above, and NASNet, described below. NASNet learns two types of modules: normal cells for feature extraction and reduction cells for down sampling input. The architecture is built by stacking these cells in a predefined arrangement. While the overall structure is fixed, the specific configurations of the cells are discovered through he reinforcement learning search strategy. The first convolutional filters and the number of motif repeats (N) are also configurable parameters. Normal Cells maintain the same feature map dimensions, while Reduction Cells reduce the height and width by half. A controller RNN predicts the cell structures by recursively selecting hidden states and operations, creating new hidden states through a series of chosen transformations. Table 2 illustrates the set of possible operations.

Table 2 displays a list of various operations commonly utilized in convolutional neural networks (CNNs), especially in the context of neural architecture search (NAS) or the development of advanced convolutional layers. Each operation plays a specific role in enhancing the network's learning capabilities. The identity operation allows the input to pass directly to the output without any modification, serving as a skip connection or when no transformation is required. The 1×7 then 7×1 convolution operation involves sequential convolutions with 1×7 and 7×1 kernels, effectively capturing

TABLE 2. A set of operation to be selected [26].

#Set of operations
1x7 then 7x1 convolution
3x3 average pooling
5x5 max pooling
1x1 convolution
3x3 depthwise-separable conv
7x7 depthwise-separable conv
1x3 then 3x1 convolution
3x3 dilated convolution
3x3 max pooling
7x7 max pooling
3x3 convolution
5x5 depthwise-separable conv

patterns in both vertical and horizontal directions, providing greater flexibility than traditional square kernels. The 3 × 3 average pooling reduces the spatial dimensions of feature maps by averaging values within a 3 × 3 window, facilitating down sampling while retaining the average intensity of the features. On the other hand, 5 × 5 max pooling selects the maximum value within each 5 × 5 window, emphasizing the most prominent features and making the model more robust to small positional changes. The 1 × 1 convolution operation is typically used to reduce or maintain the depth of feature maps, enabling dimensionality reduction without altering spatial dimensions, and is especially effective when followed by activation functions, introducing non-linearity into the model. 3 × 3 depthwise-separable convolution separates the spatial and channel dimensions into two steps—depthwise convolution (3 × 3) followed by pointwise (1 × 1) convolution—improving computational efficiency and reducing the number of parameters. 7 × 7 depthwise-separable convolution performs a similar operation but with a larger 7 × 7 spatial filter, capturing broader context and larger spatial patterns while still benefiting from the efficiency of depthwise separable convolutions. Other operations, such as 3 × 3 dilated convolution and 3 × 3 max pooling, further contribute to the model's ability to capture intricate patterns and efficiently process spatial information, each operation offering unique benefits that collectively enhance the model's performance.

E. GLOBAL AVERAGE POOLING

Compared to flattening followed by fully linked layers is a resource-intensive technique, whereas global average pooling is a more computationally efficient option. This efficiency is critical when integrating multiple transfer learning models since it considerably reduces total computing load. Global average pooling allows the combined model to maintain translation invariance while retaining important spatial information from each feature map. This versatility is especially crucial when combining transfer learning models that have been previously trained on other datasets or contexts. Global average pooling makes it easier to create a robust and efficient combined model suitable for a variety of tasks and datasets by giving a scaled representation of the input and capturing the most relevant features.

F. DENSE LAYE

The initial hidden layers consist of 512 neurons each. The output Z of each neuron is computed using the formula Z=X∗W+B, where X represents the input violent or non-violent image vector, W denotes the weight, and B represents the bias. Subsequently, the output of each hidden layer undergoes activation through the ReLU function. The resulting outputs from these hidden layers serve as inputs to the output layer. By applying the softmax function to the output layer, probabilities are computed to predict target labels. For instance, in the context of the mentioned classes, probabilities such as [1, 0] for 'not-violence' and [1, 0] for 'violence' are determined. The net input is evaluated using Equatio 1 The weight vector is indicated as 'w,' the input vector as 'x,' and the bias as 'b.' Utilizing Equation 2, softmax computation can be conducted based on Equation 1, resulting in the probabilities as described in Equations 3 and 4.

$$\mathbf{Z_i} = \frac{\mathbf{e}^{\mathbf{Z_i}}}{\sum_{\mathbf{je}} \mathbf{z_i}} \tag{1}$$

$$\mathbf{Z} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}}}{\mathbf{wi} + \mathbf{bi}} \tag{2}$$

$$\mathbf{Output} = \begin{cases} \mathbf{1}, & \mathbf{not-violence} = \mathbf{100\%} \\ \mathbf{0}, & \mathbf{violence} = 0\% \end{cases} \tag{3}$$

$$\mathbf{Output} = \begin{cases} \mathbf{0}, & \mathbf{not-violence} = \mathbf{0\%} \\ \mathbf{1}, & \mathbf{violence} = \mathbf{100\%\%} \end{cases} \tag{4}$$

G. TRANSFORMER FOR VIOLENCE AND NOT VIOLENC

This marked a major advancement in the field, as it represented the first instance of a Transformer encoder being trained on ImageNet, outperforming traditional convolutional architectures [27] The trained transformer will be applied to the Violence dataset, with all procedures detailed in Figure 5.

1) IMAGE NORMALIZATION

Normalization of images in transformer models is a vital preprocessing step that enhances the training stability and performance of the model. Standardization involves altering the pixel values so that the mean is zero and the standard deviation is one. This normalization is calculated using the formula in Equation 5:

$$\text{Normalized Pixcel value} = \frac{\text{Pixcel Value-mean}}{\text{Standard Daviation}} \tag{5}$$

In the dataset centered on violence, the mean and standard deviation are computed for each channel (R, G, B) throughout the dataset. These values, calculated using PyTorch in Python, are mean = [0.4916, 0.4818, 0.4447] and standard deviation = [0.2471, 0.2440, 0.2625], as outlined in dataset [28]

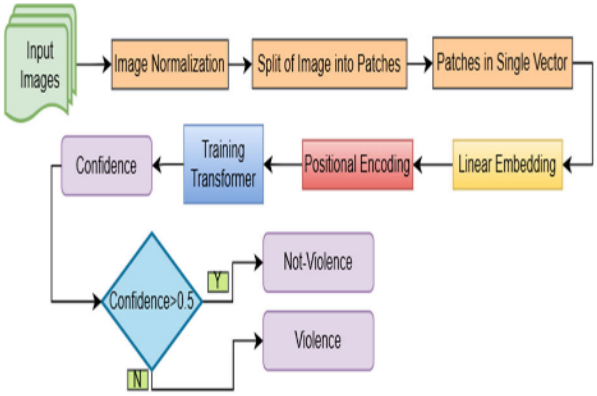

FIGURE 5. Transformation for violence and not-violence.

2) PATCHES OF IMAGE

The initial the phase in which a Vision Transformer processes an image involves segmenting it into smaller, uniform-sized patches, each representing a local area of the image.

3) PATCHES IN SINGLE VECTOR

The pixel values in each patch are Flattening into a single vector allows the model to perceive the image patches as sequential data linear Embedding. The flattened patch vectors are subsequently projected into a lower-dimensional space through trainable linear transformations, thereby reducing the data's dimensionality while retaining key features.

4) POSITIONAL ENCODING

Positional encodings are added to preserve the spatial arrangement of the patches, allowing the model t comprehend the relative positions of different patches within the image.

5) TRAINING TRANSFORMER

The combination of positional and patch embeddings makes up the input of a typical Transformer encoder. Multi-head self-attention mechanisms (MSPs), which determine attention weights to prioritize aspects of the input sequence during predictions, and multi-layer perceptron (MLP) blocks are two essential components of each of the several layers that make up this encoder. In order to properly scale and center the data within the layer and guarantee stability and effectiveness throughout training, layer normalization (LN) is conducted prior to each block. During training, an Adam optimizer adjusts the model's hyperparameters based on the loss with a learning rate of 10-4, updated in each training iteration.

6) CONFIDENCE

In the implemented model, confidence in predictions is indicated by the probability assigned to the predicted class. The class with the highest probability represents the model's most confident prediction. The MLP head in the transformer converts the processed features into a probability distribution, enabling the assessment of certainty in each prediction. If the confidence level exceeds 50%, the image is classified as Not-Violence; otherwise, it is classified as Violence.

IV. JUSTIFICATION FOR MODEL SELECTION

The selection of VGG16, MobileNetV2, NASNet, Yolov6 and a Vision Transformer model was driven by the need for a balanced approach between computational efficiency and feature extraction. While YOLOv6 models excel in object detection, they require dataset-specific tuning and struggle with contextual violence recognition. Similarly, Faster R-CNN, though highly accurate, is computationally expensive, making it less suitable for real-time surveillance. The chosen models leverage transfer learning and CNN-based feature extraction, while the Transformer model enhances global context understanding, ensuring high accuracy with lower computational costs, a crucial factor for real-time violence detection. The study strikes a balance between transfer learning and CNN-based approaches, leveraging the strengths of both while acknowledging their limitations. Future work could explore the integration of other object detection models like Faster R-CNN to enhance the system's ability to localize and classify violent incidents in real time.

V. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

Several parameters have been carefully adjusted during the model training phase to improve the effectiveness and performance of the model. Standardizing the image size to 224 by 224 pixels is an essential step in making sure the dataset is consistent. This homogeneity guarantees that the model may utilize the learnt features from pre-trained networks like VGG16 and MobileNetV2, which require input photos of a certain dimension, as well as facilitating smooth processing by convolutional neural network architectures. The batch size of 16 has been deliberately selected to achieve the best possible balance between memory usage and processing speed. With 16 photos processed at once under this arrangement, system resources are not overloaded, and the model's weights are updated based on a wide range of training instances. While bigger batch sizes can take advantage of parallel processing capabilities, smaller batch sizes can produce gradient updates that are more precise for the hardware at hand. Data augmentation strategies are critical to enhancing the model's resilience and applicability. The model's overall stability is increased, and the training process is accelerated by rescaling the picture pixel values to lie within [1, 0]. Two further augmentations that provide variation to the training dataset and faithfully capture the different scales and viewpoints of the input photographs are random shearing and zooming. Horizontal flipping significantly diversifies the dataset, reduces overfitting, and allows the model to learn more widely based properties by providing mirrored copies of the pictures. Furthermore, it has been shown that the binary classification mode can differentiate between images that depict violence and those that do not, which is significant for the specific classification issue at hand. This option allows you

TABLE 3. Train and test from dataset.

Dataset	Sets	Anomaly	Not-Anomaly
Violence Dataset	Training Set	670	1100
	Test Set	1280	600
Road Anomaly	Training Set	1242	1179
	Test Set	603	630

to customize the model to predict one of two groups, which streamlines the assessment process. Together, these adjusted parameters improve the model's overall performance and ability to successfully generalize to new data., increasing the model's accuracy and dependability in classification outputs.

B. DATASETS

Obtained from [28] the dataset consists of 4000 images depicting both 'violent' and 'non-violent' scenes. A key challenge was to devise an effective method to partition these images to cater to the model's requirements. Through meticulous examination, each image was carefully considered, resulting in the allocation of 2000 images for training and 1000 images for the testing set. Another widely used dataset is Road Anomaly Detection Dataset [29] which exclusively contains images of the Anomaly class. To obtain images for the Not-Anomaly class, images were extracted from the 'Road Vehicle Images Dataset' [30] This approach allowed the compilation of a comprehensive dataset with both Anomaly and Not-Anomaly images for the study. The distribution of training and test set data is shown in the Table 3. To evaluate the performance of the Transformer model, the training set contains 50% fewer 'Anomaly' images compared to the test set, and the test set contains 50% fewer 'Not-Anomaly' images compared to the training set. The Road Anomalies Image Dataset consists of images depicting various road irregularities and unexpected obstacles that could pose hazards to drivers. These anomalies include potholes, cracks, roadblocks, construction sites, fallen objects, debris, and accidents. The dataset is designed for anomaly detection in traffic monitoring and autonomous driving systems, helping machine learning models identify and respond to hazardous road conditions in real time. The images are captured in diverse lighting and weather conditions, ensuring robustness for real-world deployment.

C. RESULTS AND EVALUATIONS

Table 4 illustrates that the proposed model achieved a precision of 77% for the ''Not-Violence'' category and a recall of 71% for the ''Violence'' category, as highlighted by the dark color in the table. However, these metrics also indicate that there is room for improvement in the model's performance. The remaining 23% in precision for the ''Not-Violence'' category suggests that the model falsely classified some instances as ''Not-Violence.'' Similarly, the 29% gap in recall for the ''Violence'' category points to missed ''Violence'' instances.

TABLE 4. Measures from confusion matrix usingVGG16.

Class	Precision	Recall	F1 Score
Not-Violence (0)	**77%**	98%	**86%**
Violence (1)	98%	**71%**	**82%**
Average	**87%**	**84%**	**84%**
Accuracy	**84%**		

TABLE 5. Measures from Confusion Matrix using MobileNetV2.

Class	Precision	Recall	F1 Score
Not-Violence (0)	95%	**85%**	90%
Violence (1)	**86%**	96%	91%
Average	91%	90%	90%
Accuracy	90%		

TABLE 6. Measures from Confusion Matrix using YOLOv6.

Class	Precision	Recall	F1 Score
Not-Violence (0)	**90%**	92%	**91%**
Violence (1)	93%	**90%**	**92%**
Average	**92%**	**93%**	**92%**
Accuracy	**92%**		

TABLE 7. Measures from confusion matrix using ensemble-model.

Class	Precision	Recall	F1 Score
Not-Violence (0)	**92%**	98%	**94%**
Violence (1)	97%	**91%**	**94%**
Average	**94%**	**94%**	**94%**
Accuracy	**94%**		

The Table 4 includes highlighted cells indicating areas needing improvement.

Table 5 illustrates that the proposed model achieved 86% precision for the Violence category and 85% recall for the Not-Violence category, as indicated by the dark color in the table. This suggests there is still room for improvement.

Table 6 illustrates that the proposed model achieved 93% precision for the Violence category and 92% recall for the Not-Violence category, as indicated by the in the table. This suggests there is still room for improvement.

Table 7 shows that the proposed model achieved 94% accuracy, with other metrics also exceeding 90% but remaining below 95%. While the results are excellent, the potential for further improvement is indicated by the lighter colors highlighted in **Error! Not a valid bookmark self-reference** suggesting there is still room for enhancement.

The transformer for the Violence dataset was implemented in Python using torch vision transforms, utilizing the model 'google/vit-base-patch16-224-in21k'. It achieved 99% accuracy, as detailed in all the measures presented in the Table 8.

TABLE 8. Measures from confusion matrix using transformer.

Class	Precision	Recall	F1 Score
Not-Violence (0)	98%	99%	99%
Violence (1)	99%	98%	99%
Average	99%	99%	99%
Accuracy	99%		

TABLE 9. Measures from confusion matrix using transformer on road-anomaly dataset.

Class	Precision	Recall	F1 Score
Not-Violence (0)	**100%**	**96%**	**98%**
Violence (1)	**93%**	**100%**	**96%**
Average	**98%**	**97%**	**98%**
Accuracy	**98%**		

Table 9 shows that the proposed model achieved 98% accuracy using Transformer on Road-Anomaly Dataset. It achieved 98% accuracy, as detailed in all the measures presented in the Table 9.

D. EVALUATIONS

The experimental results from the VGG16, MobileNetV2, Ensemble Model, and Transformer Model are evaluated using several key metrics. These metrics provide comprehensive insights into the performance and effectiveness of each model across various aspects of classification tasks. The following metrics are utilized.

1) TRUE PREDICTION

The Transformer model outperformed the other models in predicting instances of both Not-Violence and Violence, accurately identifying 622 Not-Violence cases and 593 Violence cases. In comparison, the VGG16 model predicted 491 instances as Not-Violence and 353 as Violence, while MobileNetV2 correctly identified 423 Not-Violence and 479 Violence cases. The Ensemble model also performed well, predicting 488 instances of Not-Violence and 455 instances of Violence, indicating that the Transformer model had the highest accuracy in distinguishing between these categories as shown Figure 6.

This performance of Transformer Model is significantly better than the VGG16, MobileNetV2, and Ensemble models, as depicted in the Figure.

2) FALSE PREDICTIO

The comparative performance analysis of the models, as illustrated in Figure 7, highlights the superior accuracy of the Transformer model in minimizing false predictions. The VGG16 model produced 9 false Not-Violence predictions and 147 false Violence predictions, showing significant room for improvement in correctly identifying instances of violence.

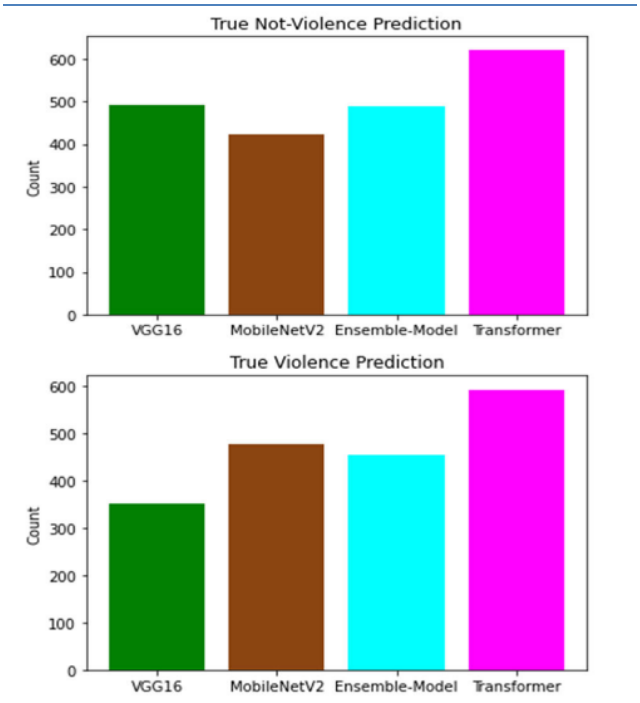

FIGURE 6. True Prediction Using VGG16, MobileNetV2,Ensemble-Model and Transformer on Violent and non-Violent Dataset.

The MobileNetV2 model, while better in predicting violence, still resulted in 77 false Not-Violence predictions and 21 false Violence predictions. The Ensemble model improved upon these results, generating 12 false Not-Violence predictions and 45 false Violence predictions. However, the Transformer model stood out by making only 8 false Not-Violence predictions and 10 false Violence predictions. This demonstrates the Transformer's effectiveness, clearly outperforming the VGG16, MobileNetV2, and Ensemble models in reducing false predictions.

3) VISUAL REPRESENTATION OF CONFUSIO MATRIX

The visual representation of the confusion matrix in Figure 8 demonstrates that the Transformer model outperforms all others in every aspect.

The experimental results indicate that, out of 1,280 road anomaly images in the dataset, 1,233 were accuratelyb identified as true anomalies, while 47 were incorrectly classified as 'Not-Anomaly'. Similarly, all 600 images classified as 'Not-Anomaly' were correctly identified. No images were falsely classified as 'Anomaly'. These findings are illustrated in Figure 9.

4) HIGHLIGHTED ISSUES FROM CONFUSION MATRIX TABLES

All highlighted issues from VGG16 in Table 4 are compared with Transformer results from Figure 10. The comparison, illustrated in the Figure 10, demonstrates that the Transformer

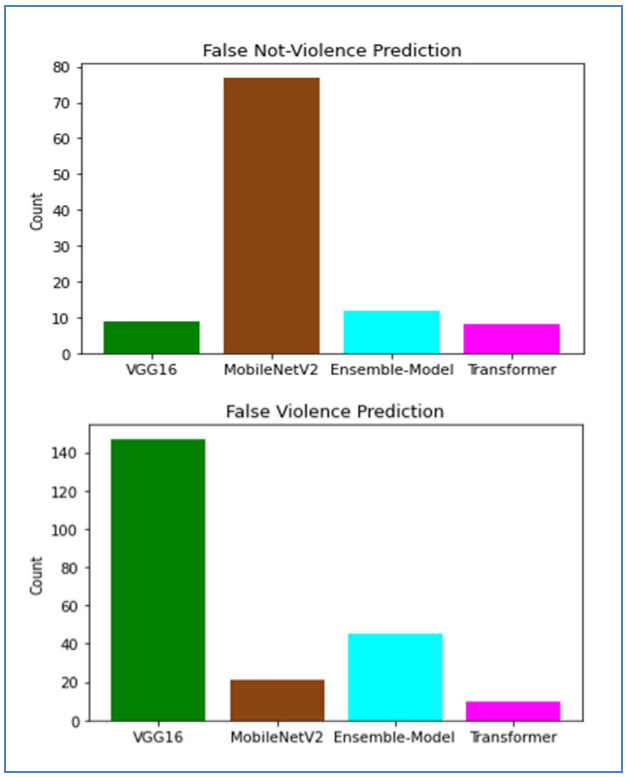

FIGURE 7. False prediction for all models.

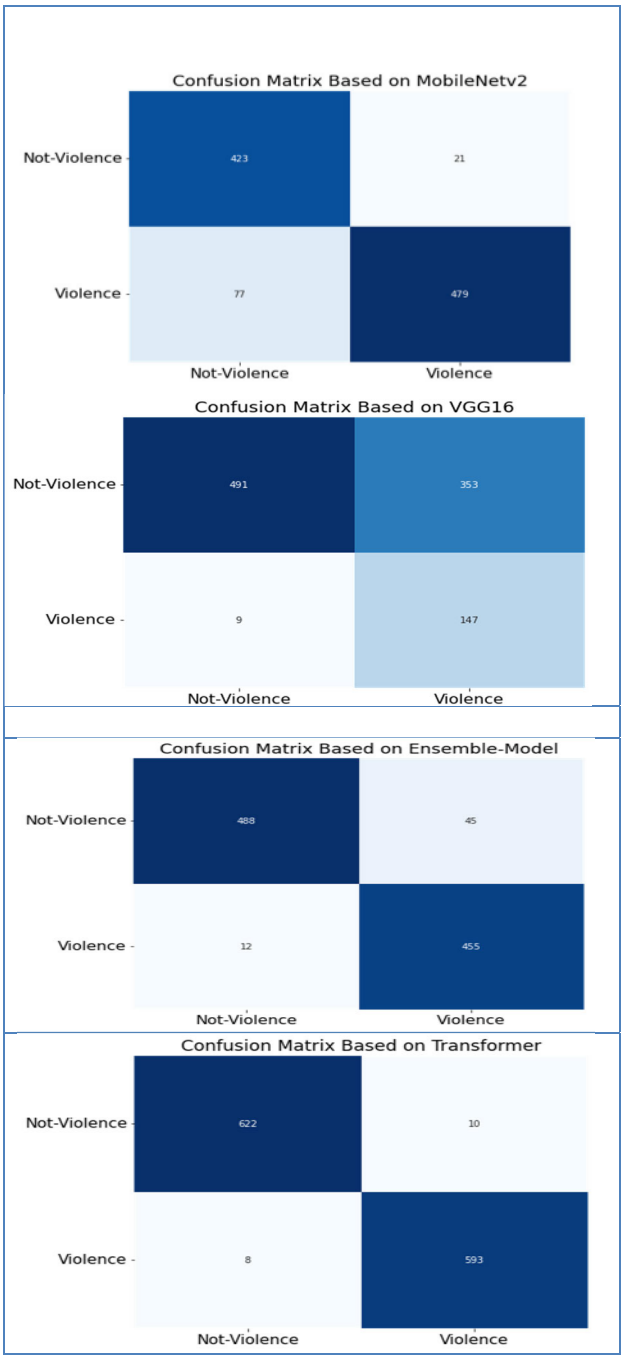

FIGURE 8. Visual representation of confusion matricesfor all models.

effectively addresses the research gaps identified in the VGG16 model.

All highlighted issues from MobileNetV2 in Table 5 are compared with Transformer results from Table 8. The comparison, illustrated in the Figure 11 demonstrates that the Transformer effectively addresses the research gaps identified in the VGG16 model.

All highlighted issues from Ensemble-Model in Table 7 shows that the proposed model achieved 94% accuracy, with other metrics also exceeding 90% but remaining below 95%. While the results are excellent, the potential for further improvement is indicated by the lighter colors highlighted in **Error! Not a valid bookmark self reference**. suggesting there is still room for enhancement.

Table 7 are compared with Transformer results from Table 8. The comparison, illustrated in the Figure 12, demonstrates that the Transformer effectively addresses the research gaps identified in the VGG16 model. demonstrates that the Transformer effectively addresses the research gaps identified in the VGG16 model.

5) ACCURACY BASED PERFORMANCE

By figuring out the proportion of accurate predictions to all input samples, this metric evaluates the model's overall prediction accuracy. A high accuracy indicates that most inputs can be correctly classified by the model.

Figure 13 provides a visual comparison of classification accuracy obtained by four different models. Each time a specific model and its related performance are color-coded. Green Bar shows the VGG16 model, which acquired an accuracy of about 84%. Brown bar corresponds to the Mobilenetv2 model, which shows a high accuracy of about 90%. Sian Bar represents the dress model, which further improves the performance with an accuracy of approximately 94%. Finally, the magenta bar highlights the transformer model, which improves others with an impressive accuracy near 99%. This picture representation clearly reflects

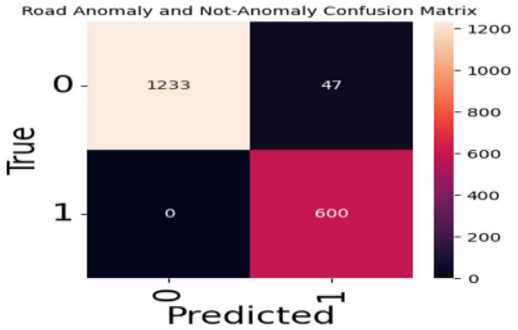

FIGURE 9. Confusion matrix of road anomalies.

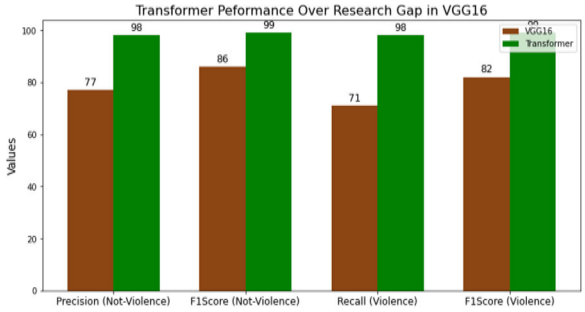

FIGURE 10. Transformer overcomes for VGG16,allissues highlighted in Table 4.

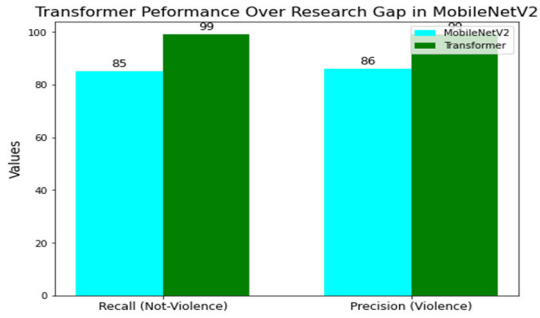

FIGURE 11. Transformer overcomes all issueshighlighted in Table 5.

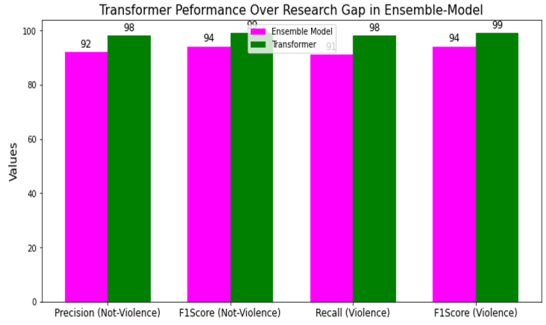

FIGURE 12. Transformer overcomes all issueshighlighted in Table 6.

progressive improvement in accuracy in the model, emerging as the most effective in transformer model four.

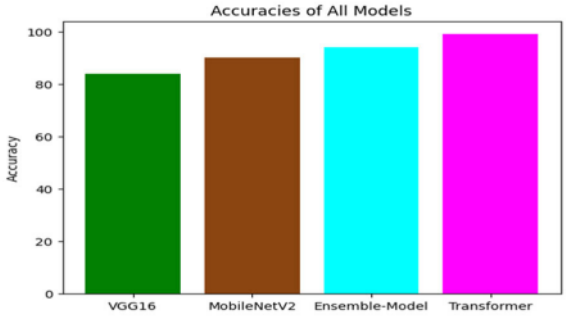

FIGURE 13. Accuracies of VGG16, MobileNetV2,ensemble-model and transformer.

6) ROC CURVE

Receiver Operating Characteristic (ROC) Curves are used to evaluate how effectively a test or set of tests balances sensitivity (the ability to correctly identify positives) and specificity (the ability to correctly identify negatives) at various cutoff points. The area under the ROC curve (AUC) measures the test's performance within a model. Figure 14 presents the ROC curve for all models on the test data. The ROC curve for the Transformer model encompasses a larger area compared to all other models, indicating that the Transformer outperforms the others in terms of effectiveness.

Figure 15 displays the receiver operating characteristics (ROC) curve for a transformer model that is assessed on a road-disbelief dataset. The orange line represents the ROC curve, which plotting the true positive rate (TPR) against the false positive rate (FPR) on various classification thresholds. The collapse blue line represents a random classifier, which provides a base line for comparison. The area under this ROC curve is the area under the Major Tech Uve Curve (AUC), which has been described as 0.98. This very high AUC value indicates that the transformer model has an excellent ability to differentiate between normal road conditions and discrepancies. The curve moves rapidly towards the topleft corner, indicating that the model receives a high true positive rate with less false positive rates. This suggests that the transformer model is highly effective in detecting road anomalies with some false alarms. The proposed model's mean absolute error is 0.025, which is a minimal value that reflects the model's excellence.

The results derived from the Road-Anomaly Dataset for the proposed model indicate its robustness and versatility. These findings demonstrate that the selected model is not only well-suited for the Road-Anomaly Dataset but also performs effectively on the violence-dataset. This suggests that the model has strong generalization capabilities and can be applied successfully to different types of datasets, confirming its reliability and adaptability in diverse contexts.

E. MEAN ABSOLUTE ERROR

Mean Absolute Error (MAE) calculates the average amount of a series of predictions' errors without taking into account

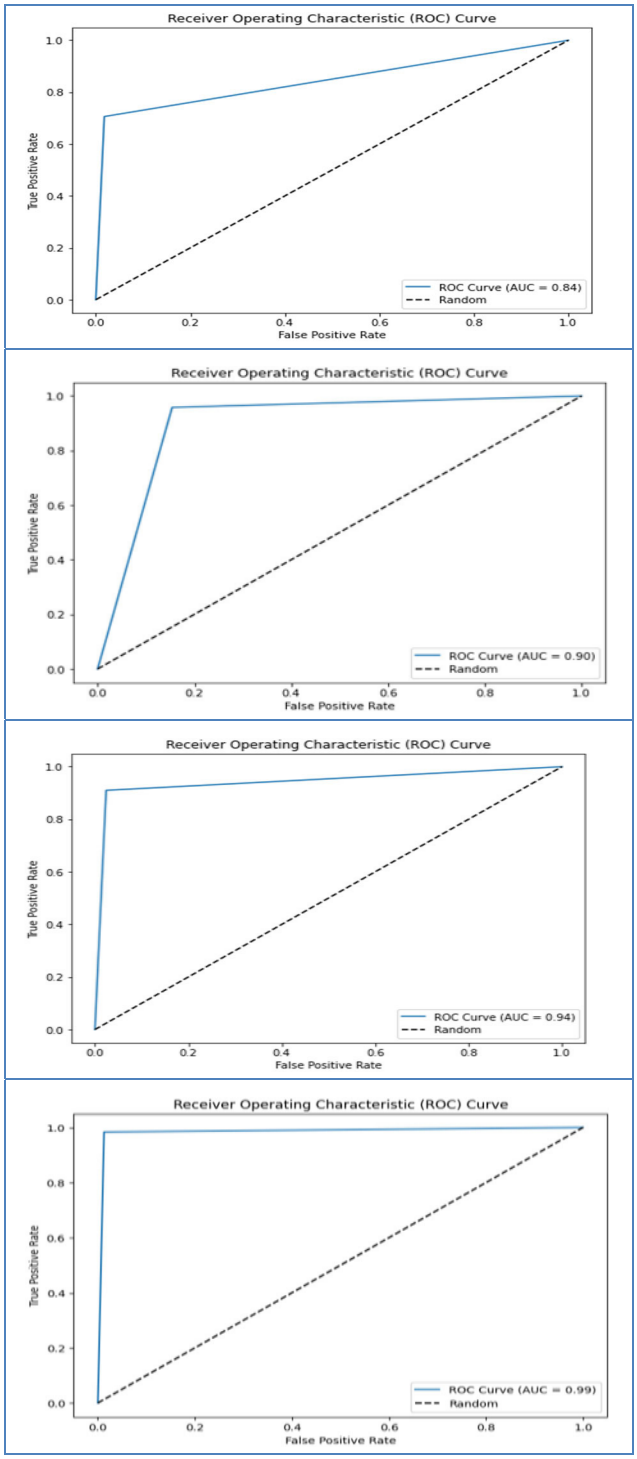

FIGURE 14. ROC Curve for VGG16, MobileNetV2,ensemble-model and transformer.

their direction. It determines the mean absolute discrepancies between expected and observed values. giving equal weight to all individual differences. Better performance is indicated by a lower MAE, which shows that the model's predictions are more in line with the actual data. The MAE values for VGG16, MobileNetV2, Ensemble-Model, and Transformer

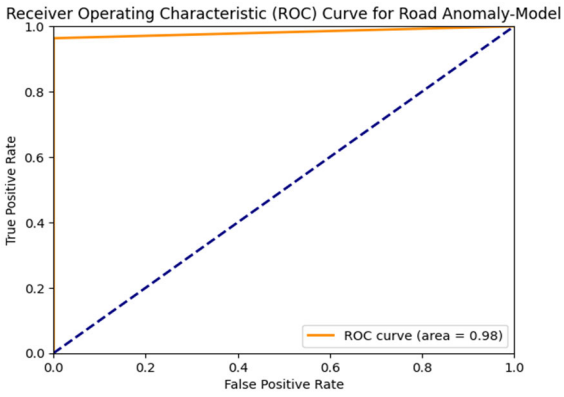

FIGURE 15. ROC curve for road-anomaly datasetusing transformer-model.

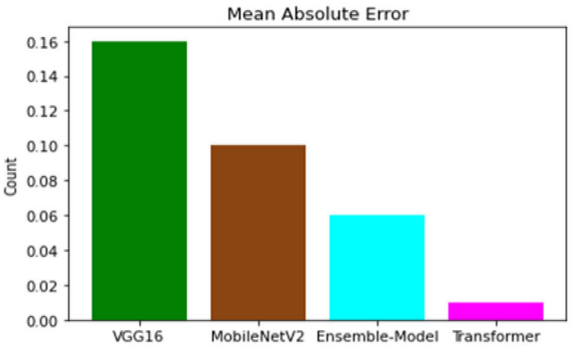

FIGURE 16. Mean absolute error for all models.

are 0.16, 0.10, 0.06, and 0.01, respectively. This minima MAE for the Transformer model demonstrates its superior performance, as illustrated in Figure 16.

1) COMPREHENSIVE EALUATION OF THE TRANSFORMER MODE'S SUPERIOR PERFORMANCE

The Transformer model outperformed other architectures, including VGG16, MobileNetV2, and the Ensemble Model, based on key evaluation metrics such as the confusion matrix, training loss behavior, ROC curves, and Mean Absolute Error (MAE). These metrics highlight the model's superior ability to distinguish between violent and non-violent incidents, making it the most reliable choice for smart surveillance systems.

One of the most critical aspects of model evaluation is the confusion matrix analysis. The Transformer model exhibited the lowest False Positive Rate (FPR) and False Negative Rate (FNR) compared to other models. For instance, VGG16 misclassified 583 non-violent images as violent, while the Transformer model only misclassified 10. Similarly, in detecting violent events, MobileNetV2 misclassified 77 cases, whereas the Transformer model misclassified just 8. These results indicate that the Transformer model achieves higher precision and recall, minimizing false alarms and

missed detections—crucial factors for real-time security applications.

Another indicator of superior performance is training loss stability. Unlike other models, which exhibit oscillating loss values, the Transformer model shows a steady decrease in training loss, ensuring a more stable and efficient learning process. This behavior indicates that the Transformer effectively extracts meaningful patterns from data without overfitting, allowing it to generalize well across different surveillance scenarios.

The ROC curve and AUC score further validate the Transformer's dominance. The Transformer model achieved a significantly higher AUC than VGG16, MobileNetV2, and the Ensemble Model, meaning it has superior sensitivity and specificity. A higher AUC implies that the Transformer can more accurately classify violent and non-violent events, which is essential for reducing response time in security interventions.

Lastly, the Mean Absolute Error (MAE) was notably lower for the Transformer model compared to the other architectures. A lower MAE signifies that the model's predictions are more accurate and consistent with real-world observations. This metric confirms the Transformer's ability to deliver precise and reliable results, ensuring its effectiveness for real-time applications.

2) EXPLORING TRANSFORMER MODEL

Since the transformer model was chosen as the most suitable model for the proposed work, it is worthwhile to study the performance characteristics of this model and its deficiencies, if any. The Transformer model misclassified the following images as violence which were not violent at all: 8 images. In Table 10, we inspect examples where non -violent images are incorrectly performed as violence. This abortion can arise from various factors, such as a bad light condition, where it makes it difficult to differentiate between violent and non -violent behaviors for low contrast models. Additionally, the anxiety (eg, partially obstructed objects or individuals) can be incomplete or vague facility, resulting in incorrect predictions. To reduce these issues, data enlargement techniques such as brightness, contrast, and noise can help the model become stronger for different lighting conditions. In addition, applying techniques such as image cropping and flipping can address the obstacle by providing models with more diverse ideas of the same scenes. Incarnating temporary information from adjacent frames in video data can improve classification accuracy by providing additional references and reducing the possibility of separate abortion.

In Table 11, we look at cases where violence is done wrong as non -violence. These errors can be attributed to factors such as rapid movement of view or lack of different characteristics that clearly indicate violent activity. In these conditions, the model may fail to detect subtle or transient violent behaviors, especially when actions are fleeting or rapid speed. To measure it, incorporating temporary information from the surrounding frame will allow the model to occupy the dynamics of violent events that may not be clear in the same frame. Additionally, employing advanced techniques such as temporal smoothing or motion detection can help the model better understand the context of the scene, which shows more accurate violence.

Here, highlight two particular instances of misclassification for additional transparency and to help readers better understand the discussion in 10 and 11. In the first instance, picture 173 was mistakenly predicted by the model to be Violence when it was really tagged as Non-Violence. In the second instance, image-1156 was incorrectly identified as non-violent when it was first classed as violent. As seen in Figure 17, these examples highlight the model's shortcomings in differentiating between some complicated cases.

TABLE 10. Not-violence detected as violence using transformer.

Image Name	Actual Label	Predicted Label	Confidence
Image-150	Non-Violence	Violence	0.6830
Image-173	Non-Violence	Violence	0.9943
Image-232	Non-Violence	Violence	0.5270
Image-248	Non-Violence	Violence	0.9971
Image-255	Non-Violence	Violence	0.7540
Image-303	Non-Violence	Violence	0.9898
Image-478	Non-Violence	Violence	0.9852
Image-528	Non-Violence	Violence	0.6259

TABLE 11. Violence detected as not-violence using transformer.

Image Name	Actual Label	Predicted Label	Confidence
Image-855	Violence	Non-Violence	0.5531
Image-942	Violence	Non-Violence	0.5263
Image-986	Violence	Non-Violence	0.9974
Image-1020	Violence	Non-Violence	0.9898
Image-1156	Violence	Non-Violence	0.9568
Image-1186	Violence	Non-Violence	0.8364
Image-1212	Violence	Non-Violence	0.9767
Image-1227	Violence	Non-Violence	0.9944

FIGURE 17. Sample from wrong predicted images.

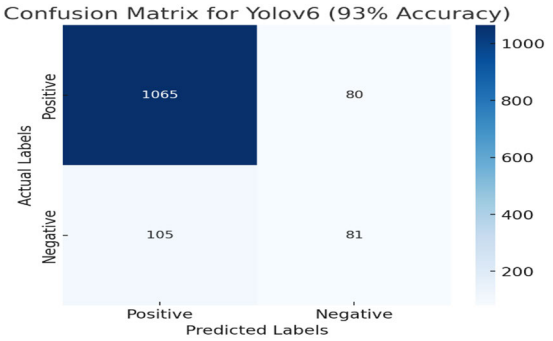

FIGURE 18. Visual representation of confusion matricesfor Yolov6 model.

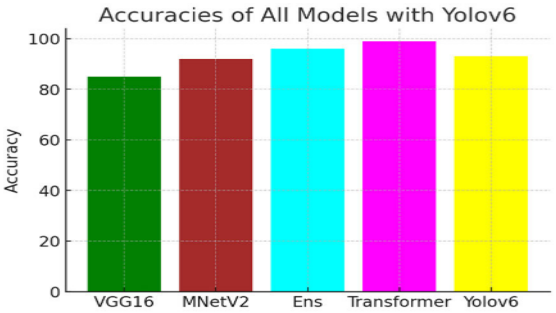

FIGURE 19. Comparison of all Models with Yolov6.

We may better understand the mechanisms behind these mistakes and provide insights into possible areas for model development by examining these misclassifications.

3) YOLO BASED PERFORMANCE

We implemented YOLOv6 due to its high speed and balanced accuracy, which is essential for detecting violence in real time environment. Its design is both efficient and lightweight, allowing it to quickly diagnose fast-moving actions and small gestures that might indicate violent behavior. This means it has the capability to process video streams without requiring massive computational power, making it a practical choice for systems running on everyday hardware. YOLOv6 has been used for object recognition to distinguish between aggression and non-violence in addition to the previously used classification models—VGG16, MobileNetV2, Ensemble Model, and Transformer. YOLO v6 concentrates on identifying certain objects connected to violent acts, as opposed to classification algorithms that examine the full image. But as Table 6 demonstrates, its total accuracy is just 93%, which is below the classification models' performance. This restriction is further demonstrated in Figure 18, which graphically displays the results of YOLOv6 object detection.

The ROC (Receiver Operating Characteristic) curve and Precision-Recall curve for the YOLO model did not yield satisfactory results, indicating its limitations in effectively distinguishing between violence and non-violence. As shown in Figure 19, the ROC curve suggests a lower true positive rate compared to false positives, reflecting poor discrimination capability. Similarly, the Precision-Recall curve highlights inconsistencies in model predictions, with lower precision and recall values, further emphasizing the challenges YOLO faces in this classification task. These results reinforce that YOLO is not well-suited for this particular application, as its performance is significantly lower than the classification models in terms of both detection accuracy and reliability.

Additionally, as seen in the comparative performance graph in Figure 19, YOLOv6 accuracy is lower than that of the Ensemble Model succeeding at 94%, and the Transformer model excelling at 99%. These results reinforce that while YOLO is effective for object detection, it is not the optimal choice for image classification in this scenario.

4) ABLATION STUDY

In Work [31], the authors utilize cascaded Temporal Shift Modules (TSMs) inspired by Long Short-Term Memory (LSTM) networks to enhance long-term feature extraction, achieving an accuracy of 98% in violence behavior recognition. In contrast, the proposed work employs a Transformer architecture with attention hyperparameters, resulting in a higher accuracy of 99%. This comparison highlights that the Transformer-based approach outperforms the TSM-based method, demonstrating superior capability in modeling long-range dependencies.

In Work [32], which uses the combination of MobileNet-V2 and the Temporal Shift Module, the proposed network achieves the accuracy of 98%. On the other hand, the proposed work utilizes the Transformer architecture with some

important hyperparameters for attention and has an accuracy of 99%. As such, this comparison illustrates that the Transformer-based approach offers more efficient handling of complicated temporal patterns to achieve higher results. Work [33] has employed an advanced IIOT system to detect violence in industrial premises with 98 % accuracy, however the disadvantage of limited coverage of indoor activities, high computational complexity, and high latency, for real time basically. Thus, the proposed work with the help of a neural network that is Transformer architecture, having attention hyperparameters selection results in the accuracy of 99%. This leads to the speculations that the Transformer approach could offer superior performance in terms of dealing with time orders and real-time processing. In [34], a hybrid model using U-Net as the spatial component and MobileNet V2 along with LSTM for the temporal component has been applied with an accuracy of 95%. The proposed work is built with Transformer architecture and has other hyperparameters related to the attention and it has an accuracy of 99%. This implies that the Transformer model may give better results in terms of the ability to learn temporal dependencies stronger relative to the hybrid U-Net and LSTM approach developed in work [35].

Work [36]presents an Unsupervised Domain Adaptation method for video violence detection having a accuracy of 98% that can solve the problem of domain shift between the source and target video clips. The work introduced in the paper that employs Transformer model with selected attention hyperparameters gets a higher level of accuracy of 99%. However, this implies that Transformer model might help achieve better performance in terms of the feature interactions as well as in domain adaptation in contrast to the method described in Work features VD-Net, an AI-based framework that uses ST-TCN blocks and bottleneck layers to detect violent behavior in various settings, achieving an accuracy of 98%. In comparison, the proposed work, based on a Transformer architecture with attention hyperparameters, achieves a higher accuracy of 99%. This indicates that the Transformer model may provide better performance in understanding complex temporal dependencies, enhancing the effectiveness of violence detection. The comparison of the proposed model with similar studies is presented in the Table 3 and visual representation is found in Figure 20.

VI. SUGGESTION ABOUT REAL-TIME IMPLEMENTATION

The real-time performance of the proposed Transformer model is a key consideration for its implementation in public safety systems. Since standard surveillance cameras operate at 30 frames per second (FPS), processing every frame would be computationally expensive and could introduce latency. A more efficient approach is to analyze keyframes at an interval of 250–500 milliseconds (i.e., every 8 to 15 frames at 30 FPS), ensuring timely detection while optimizing computational resources. The Transformer model, when optimized using TensorRT or ONNX Runtime on edge AI chips such as NVIDIA Jetson Xavier or Google Coral TPU, can achieve an inference time of less than 100 milliseconds per frame, enabling near-instantaneous violence detection. Once a frame is classified as violent, the system can buffer and analyze the subsequent 2–3 frames to validate the event before triggering an alert. If violent activity is consistently detected, the system can automatically send notifications to security personnel via an embedded AI module (e.g., an FPGA or Jetson Nano-based deployment). This strategy ensures an optimal balance between detection speed and computational efficiency, making the Transformer model highly effective for real-time surveillance applications.

TABLE 12. Comparison of proposed model with benchmarks.

Models	Accuracy
Census Transform Histogram [15]	82%
Machine Learning Models [17]	71%
Two-TSM Method [31]	98%
TSM-MobileNet [32]	97%
Darknet [33]	98%
MobileNet-TSM [37]	97%
U-Net [34]	95%
Vit-Large [35]	98%
VD-Net [36]	98%
M1- Proposed Model: Transformer (Real-life-violence-situation)	**99%**
M2- Proposed Model: Transformer ('Road Anomalies Image Dataset)	**98%**

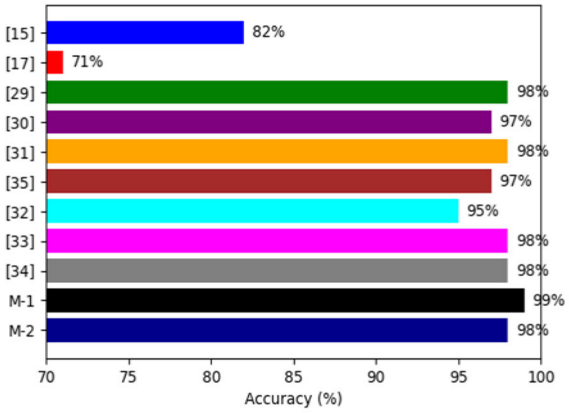

FIGURE 20. Comparative analysis of proposed modelswith existing models.

VII. CONCLUSION

This study focuses on detecting violent situations to enhance public security. It investigates the use of AI for identifying violent scenes in video frames, showcasing notable advancements in both the speed and accuracy of violence detection. The research involves the development and rigorous testing of four deep learning models—VGG16, MobileNetV2, an ensemble model, and a Transformer model—using publicly available datasets with images of both violent and

non-violent content. Among these models, the Transformer model achieved the highest accuracy at 99%, significantly outperforming the others. It also excelled in other key metrics, including precision, recall, and F1 score, making it the most effective for real-time violence detection. The ensemble model also demonstrated strong performance, particularly in precision and recall, though it did not match the overall accuracy of the Transformer model. The practical applications of these findings are significant, with the Transformer model proving highly effective when deployed in smart cameras for public safety and on smartphones to filter violent content, creating safer environments both physically and virtually. Furthermore, the Transformer model's robustness was confirmed by its strong performance on the Road-Anomaly dataset, indicating its versatility and reliability across different contexts. These results suggest that the Transformer model, alongside the ensemble model, is the most suitable for implementation in violence detection and classification tasks, offering a powerful tool for enhancing public safety and protecting vulnerable populations. As the proposed model achieved the best results, there are still several areas that offer potential directions for future work. The current research utilized VGG16, MobileNetV2, and google/vit-base-patch16-224-in21k models, and tested them on datasets containing 4,000 and 3,650 images. Future studies could explore expanding this range and applying the model to additional datasets to enhance its robustness and generalization capabilities. Additionally, designing and testing user-centered interfaces would be beneficial, allowing non-technical users, such as parents or security personnel, to interact more effectively with AI-based violence detection systems. This would improve both accessibility and practical application, making the technology more user-friendly and adaptable to real-world scenarios.

Compared to recent transformer-based models such as ViVits and videos swin transformers, our job also extends the application of transformer models to detect video-based violence. While models such as ViVit are sewn to video data by processing temporary information in many frames, our approach focuses on analyzing the keyframes efficiently at an interval of 250–500 milliseconds to reduce computational costs while maintaining real-time performance. Transformer model, NVIDIA jetson xavier or Google Coral TPU -like Edge AI Chips using Tensors or ONNX runtime using the runtime, can achieve an estimated time under 100 milliseconds per frame. This allows rapid violence to detect, validate the incident in the later frame before sending an alert to the system. This method creates an ideal balance between computational efficiency and address accuracy, enhancing real -time monitoring systems for public safety.

AUTHORS CONTRIBUTIONS

All authors contributed to this research article's design, analysis, experiments, writing, and revisions. All authors revised and approved the final version of this manuscript.

CONFLICT OF INTEREST

The authors declare no competing financial and non-financial interests.

REFERENCES

[1] M. A. Jensen, A. Atwell Seate, and P. A. James, "Radicalization to violence: A pathway approach to studying extremism," *Terrorism Political Violence*, vol. 32, no. 5, pp. 1067–1090, Jul. 2020, doi: 10.1080/09546553.2018.1442330.

[2] F. Thijs, E. Rodermond, E. R. Kleemans, and S. G. A. van de Weijer, "Violent and nonviolent terrorist suspects: A comparative analysis based on data from The Netherlands," *Eur. J. Criminal Policy Res.*, vol. 30, no. 1, pp. 63–83, Mar. 2024, doi: 10.1007/s10610-022-09523-9.

[3] I. Serrano Gracia, O. Deniz Suarez, G. Bueno Garcia, and T.-K. Kim, "Fast fight detection," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0120448, doi: 10.1371/journal.pone.0120448.

[4] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight recognition in video using Hough forests and 2D convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4787–4797, Oct. 2018, doi: 10.1109/TIP.2018.2845742.

[5] A. Mumtaz, A. B. Sargano, and Z. Habib, "Violence detection in surveillance videos with deep network using transfer learning," in *Proc. 2nd Eur. Conf. Electr. Eng. Comput. Sci. (EECS)*, Dec. 2018, pp. 558–563, doi: 10.1109/EECS.2018.00109.

[6] B. Cao, H. Xia, and Z. Liu, "A video abnormal behavior recognition algorithm based on deep learning," in *Proc. IEEE 4th Adv. Inf. Manag., Communicates, Electron. Autom. Control Conf. (IMCEC)*, Jun. 2021, pp. 755–759, doi: 10.1109/IMCEC51613.2021.9482114.

[7] S. Mohammadi, H. Kiani, A. Perina, and V. Murino, "Violence detection in crowded scenes using substantial derivative," in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2015, pp. 1–6, doi: 10.1109/AVSS.2015.7301787.

[8] A. F. Alqahtani and M. Ilyas, "A machine learning ensemble model for the detection of cyberbullying," *Int. J. Artif. Intell. Appl.*, vol. 15, no. 1, pp. 115–129, Jan. 2024, doi: 10.5121/ijaia.2024.15108.

[9] S. A. Sumon, R. Goni, N. B. Hashem, T. Shahria, and R. M. Rahman, "Violence detection by pretrained modules with different deep learning approaches," *Vietnam J. Comput. Sci.*, vol. 7, no. 1, pp. 19–40, Feb. 2020, doi: 10.1142/s2196888820500013.

[10] Y. Alotaibi and M. Ilyas, "Ensemble-learning framework for intrusion detection to enhance Internet of Things' devices security," *Sensors*, vol. 23, no. 12, p. 5568, Jun. 2023, doi: 10.3390/s23125568.

[11] A. Inbavalli, T. Jarshini, and M. Muralikrishnaa, "Efficient aggressive behaviour detection and alert system employing deep learning techniques," in *Proc. Int. Conf. Autom. Comput. (AUTOCOM)*, Mar. 2024, pp. 404–410, doi: 10.1109/AUTOCOM60220.2024.10486157.

[12] K. P. Saranyanath, W. Shi, and J. P. Corriveau, "Cyberbullying detection using ensemble method," in *Proc. CS IT Conf.*, 2022, vol. 12, no. 15, pp. 1–20, doi: 10.5121/csit.2022.121507.

[13] A. Ali and A. M. Syed, "Cyberbullying detection using machine learning," *Pakistan J. Eng. Technol.*, vol. 3, no. 2, pp. 45–50, 2020.

[14] A. F. Alqahtani and M. Ilyas, "An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying," *Mach. Learn. Knowl. Extraction*, vol. 6, no. 1, pp. 156–170, Jan. 2024, doi: 10.3390/make6010009.

[15] F. de Souza and H. Pedrini, "Detection of violent events in video sequences based on census transform histogram," in *Proc. 30th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2017, pp. 323–329, doi: 10.1109/SIBGRAPI.2017.49.

[16] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, "Cyberbullying detection: An ensemble based machine learning approach," in *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV)*, Feb. 2021, pp. 710–715, doi: 10.1109/ICICV50876.2021.9388499.

[17] M. N. Dharani, M. P. Anbumani, and M. E. Sivakumar. (2022). *Cyber Bulling Detection in Chat Application*. [Online]. Available: www.kaggle.com

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Adv. Neural Inf. Process Syst.*, 2017, pp. 1–22.

[19] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," 2021, *arXiv:2103.15691*.

[20] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.
[21] S. M. Hammami and M. Alhammami, "Vision-based system model for detecting violence against children," *MethodsX*, vol. 7, Jan. 2020, Art. no. 100744.
[22] K. Le. (2021). *An Overview of VGG16 and NiN Models*. [Online]. Available: https://medium.com/mlearning-ai/an-overview-of-vgg16-and-nin-mode
[23] G. Rohini, "Everything you need to know about VGG16," *Great Learning*, 2021. [Online]. Available: https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918
[24] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
[25] M. Horvat, L. Jelečević, and G. Gledec, "Comparative analysis of YOLOv5 and YOLOv6 models performance for object classification on open infrastructure: Insights and recommendations," in *Proc. 34th Central Eur. Conf. Inf. Intell. Syst.*, Zagreb, Croatia, pp. 317–324.
[26] SikHo Tsang. (2019). *Review: NASNet—Neural Architecture Search Network (Image Classification)*. Medium. Accessed: Jan. 15, 2024. [Online]. Available: https://sh-tsang.medium.com/review-nasnet-neural-architecture-search-network-image-classification-23139ea0425d
[27] E. Blog. (2023). *Introduction To Vision Transformers (ViT), Encoder*. [Online]. Available: https://encord.com/blog/vision-transformers/
[28] A. Raja. (2023). *Violence Vs. Non-Violence: 11K Images Dataset*. [Online]. Available: https://www.kaggle.com/datasets/abdulmananraja/real-life-violence-situations
[29] *Road Anomalies, Instance Segmentation*. Accessed: Jan. 20, 2024. [Online]. Available: https://universe.roboflow.com/qu-xlrpp/road-anomalies-li62k/dataset/5
[30] A. Yeafi. *Road Vehicle Images Dataset*. Kaggle. Accessed: Jun. 1, 2024. [Online]. Available: https://www.kaggle.com/datasets/ashfakyeafi/road-vehicle-images-dataset
[31] Q. Liang, Y. Li, B. Chen, and K. Yang, "Violence behavior recognition of two-cascade temporal shift module with attention mechanism," *J. Electron. Imag.*, vol. 30, no. 4, pp. 1–13, Jul. 2021, doi: 10.1117/1.jei.30.4.043009.
[32] W. Wang, S. Dong, K. Zou, and W. Li, "A lightweight network for violence detection," in *Proc. 5th Int. Conf. Image Graph. Process. (ICIGP)*, Jan. 2022, pp. 15–21.
[33] F. U. M. Ullah, M. S. Obaidat, K. Muhammad, A. Ullah, S. W. Baik, F. Cuzzolin, J. J. P. C. Rodrigues, and V. H. C. de Albuquerque, "An intelligent system for complex violence pattern analysis and detection," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10400–10422, Dec. 2022, doi: 10.1002/int.22537.
[34] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, p. 2216, Mar. 2022.
[35] L. Ciampi, C. Santiago, J. P. Costeira, F. Falchi, C. Gennaro, and G. Amato, "Unsupervised domain adaptation for video violence detection in the wild," in *Proc. IMPROVE*, Jan. 2023, pp. 37–46.
[36] M. Khan, A. E. Saddik, W. Gueaieb, G. De Masi, and F. Karray, "VD-net: An edge vision-based surveillance system for violence detection," *IEEE Access*, vol. 12, pp. 43796–43808, 2024.
[37] Y. Zhang, Y. Li, and S. Guo, "Lightweight mobile network for real-time violence recognition," *PLoS ONE*, vol. 17, no. 10, Oct. 2022, Art. no. e0276939.

ABDULRAHMAN ALSHALAWI received the B.S. degree in computer engineering from Taif University, in 2012, and the M.S. degree from Tennessee State University, Nashville, TN, USA, in 2016. His research interests include machine learning, data science, and computer vision.

WADOOD ABDUL (Member, IEEE) received the Ph.D. degree in signal and image processing from the University of Poitiers, France, in 2011. Currently, he is a Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University. His research interests include multimedia security, biometrics, agriculture applications, privacy, medical image processing, and video understanding, where he is working on several externally funded research projects. He has published over 100 papers in well-reputed conferences and journals. He received the Best Faculty Award from the College of Computer and Information Sciences, King Saud University, in 2017.

GHULAM MUHAMMAD (Senior Member, IEEE) received the B.S. degree in computer science and engineering from Bangladesh University of Engineering and Technology, in 1997, and the M.S. and Ph.D. degrees in electronic and information engineering from Toyohashi University and Technology, Japan, in 2003 and 2006, respectively. He is currently a Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He has authored and co-authored more than 300 publications, including IEEE/ACM/Springer/Elsevier journals, and flagship conference papers. He owns three U.S. patents. He has supervised more than 15 Ph.D. and Master's Theses. He is involved in many research projects as a principal investigator and a co-principal investigator. His research interests include signal processing, machine learning, IoTs, medical signal and image analysis, AI, and biometrics. He was a recipient of Japan Society for Promotion and Science (JSPS) Fellowship from the Ministry of Education, Culture, Sports, Science and Technology, Japan.