Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.072655

ARTICLE

# Multi-CNN Fusion Framework for Predictive Violence Detection in Animated Media

**Tahira Khalil**[1], **Sadeeq Jan**[2,*], **Rania M. Ghoniem**[3] **and Muhammad Imran Khan Khalil**[1]

[1]Department of Computer Science & Information Technology, University of Engineering & Technology, Peshawar, 25000, Pakistan
[2]National Center for Cyber Security, University of Engineering & Technology, Peshawar, 25000, Pakistan
[3]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

*Corresponding Author: Sadeeq Jan. Email: sadeeqjan@uetpeshawar.edu.pk

Received: 01 September 2025; Accepted: 06 November 2025; Published: 09 December 2025

**ABSTRACT:** The contemporary era is characterized by rapid technological advancements, particularly in the fields of communication and multimedia. Digital media has significantly influenced the daily lives of individuals of all ages. One of the emerging domains in digital media is the creation of cartoons and animated videos. The accessibility of the internet has led to a surge in the consumption of cartoons among young children, presenting challenges in monitoring and controlling the content they view. The prevalence of cartoon videos containing potentially violent scenes has raised concerns regarding their impact, especially on young and impressionable minds. This article contributes to the growing concerns about the impact of animated media on children's mental health and offers solutions to help mitigate these effects. To address this issue, an intelligent, multi–CNN fusion framework is proposed for detecting and predicting violent content in upcoming frames of animated videos. The framework integrates probabilistic and deep learning methodologies by leveraging a combination of visual and temporal features for violence prediction in future scenes. Two specific convolutional neural network classifiers i.e., VGG16 and ResNet18 are employed to classify scenes from animated content as violent or non-violent. To enhance decision robustness, this study introduces a fusion strategy based on weighted averaging, combining the outputs of both Convolutional Neural Networks (CNNs) into a single decision stream. The resulting classifications are subsequently fed into Naive Bayes classifier, which analyzes sequential patterns to forecast violence in future scenes. The experimental findings demonstrate that the proposed framework achieved predictive accuracy of 92.84%, highlighting its effectiveness for intelligent content moderation. These results underscore the potential of intelligent data fusion techniques in enhancing the reliability and robustness of automated violence detection systems in animated content. This framework offers a promising solution for safeguarding young audiences by enabling proactive and accurate moderation of animated videos.

**KEYWORDS:** Violence prediction; multi-model fusion; cartoon videos; residual network (ResNet); visual geometry group (VGG); CNN

## 1 Introduction

The world of entertainment and media usage has undergone a significant transformation in recent years, mainly driven by advances in communication and transmission technologies. The rise of broadcast media and the internet has fundamentally changed how people access and engage with media content. This change is especially noticeable in the realm of cartoon videos, a genre that continues to captivate children and young audiences. One crucial aspect of children's interaction with cartoons is the influence these animations have on their language development [1]. Additionally, the distinctive clothing and styles of animated characters can inspire children's personal fashion choices, highlighting the broad influence cartoons have on behavior and self-expression [2].

Cartoons can be both entertaining and informative, but some of the content they present distorts society, with detrimental effects on kids' subconscious perceptions [3]. Whether the impacts are beneficial or harmful depends primarily on the child's developmental stage; thus, it's critical to recognize and control content that can negatively affect behavior [4]. Cartoon violence is a developing problem since it normalizes aggressive conduct, which might have an impact on children's psychosocial development [5]. Cartoon characters frequently have strong, consistent behavioral identities, with some displaying more violent behaviors in contrast to real-world videos. This allows predicting a scene's mood based on specific characters and objects [6]. Although cartoons are primarily created for entertainment and education, violent scenes may still appear, potentially causing distress and adverse behavioral outcomes. Understanding and predicting such violence is therefore crucial for parents, educators, and content creators [7]. Providing accurate, automated age-appropriate content filtering can help platforms classify animated content more effectively for compliance and suitability—the unique features of cartoons present challenges that require innovative solutions. CNN-based future scene prediction techniques offer flexibility and utility in interactive media and animation by respecting cartoons' narrative structure while anticipating upcoming scenes. Using custom datasets that capture cartoons' distinctive characteristics improves model accuracy in detecting violence. This approach supports a thorough strategy for identifying violent content, enhancing classification, and protecting young viewers. Prior research has analyzed violence in cartoons through content analysis of visual and narrative elements, forming a foundation for this study's focus.

The proposed work aims to predict future violent scenes in animated cartoons by analyzing patterns in historical data. ResNet-18 and VGG16 were chosen as the main classifiers for violence detection due to their demonstrated ability to extract visual features from stylized, low-texture animated scenes. Cartoon videos typically feature simpler textures, solid colors, and exaggerated shapes, making them well-suited for CNNs designed to recognize hierarchical patterns. Both classifiers were trained and evaluated on a customized dataset, and their results were combined using a weighted averaging fusion technique. This method was selected over alternatives, such as majority voting or stacking, for its simplicity and efficiency. Since VGG16 and ResNet-18 may produce differing confidence levels depending on visual style or character complexity, weighted averaging allows assigning greater influence to the more reliable model in each context, improving overall accuracy and robustness. The Naive Bayes classifier was employed to model temporal dynamics by estimating the probability of violence in future scenes based on patterns in previous frames. This interpretable, straightforward approach supports practical temporal inference, making it particularly suitable for real-time cartoon video analysis, where quick, accurate decisions are essential. The principal contribution of this research is the prediction of future violent scenes in animated videos, which distinguishes it from existing studies that primarily focus on real-world video violence detection. Additionally, the study introduces a customized dataset of animated clips specifically curated to train models for effective violence detection in cartoons, addressing the unique challenges posed by animated media.

The remainder of this paper is structured as follows: Section 2 outlines relevant prior research, while Section 3 discusses the problem statement. Section 4 explains the proposed framework. Section 5 introduces the dataset, followed by the presentation of results. Finally, Section 6 describes the conclusion and potential avenues for future research.

## 2 Related Work

Cartoons are a popular form of entertainment among children and play a vital role in their personality development by amusing them within a healthy environment. However, exposure to violent cartoons has been linked to desensitization and altered perceptions of violence, which raises concerns about their impact on young viewers [8]. Significant research has focused on violence detection in real-world videos. Authors in [9] introduced a three-stage framework using saliency features to identify violent scenes, enabling the creation of violence-free content for children. In [10], the authors examined movie scripts to predict violence ratings, showing that textual cues can indicate violent content even before production. Ref. [11] proposed their model for violence detection using video surveillance techniques. Despite these advances, the visual style and motion characteristics of animated content differ significantly from those of natural settings; hence, many models may not translate well to animated content. Furthermore, temporal modeling techniques frequently entail intricate architectures that might not be ideal for animation analysis.

Deep learning has played a pivotal role in advancing image and video analysis for detecting violence. To detect violence in real-world videos, Ref. [12] suggests a model that combines recurrent and convolutional neural networks to extract temporal and spatial features, thereby improving detection accuracy by effectively capturing motion patterns and frame-level data. Similarly, Ref. [13] discussed the applications of deep learning and neural networks, making it a valuable contribution to image processing. In [14], the authors examined a comprehensive survey of techniques for filtering inappropriate video content across platforms, covering both traditional and deep learning methods. They emphasize the need for effective filtering to protect viewers, especially children, and discuss the limitations of current approaches while suggesting future research directions. Other architectures, including inception-ResNet v2 [15] and MobileNetV2–ConvLSTM [16], have shown high accuracy in real-time violence detection. At the same time, graph convolutional networks have been used to cluster violent scenes temporally with enhanced reliability [17]. Models such as ResNet-50, VGG16, MobileNetV2, and EfficientNet have been successfully applied to character recognition and emotion classification in animated images [18,19]. End-to-end CNNs have outperformed traditional feature-based methods in detecting inappropriate cartoon content, supported by domain-specific datasets [19]. Hybrid models incorporating attention mechanisms and transformers have further improved the accuracy of real-time violence detection [20]. Multimodal approaches that combine visual and textual data have been explored to provide more comprehensive content classification [21]. However, challenges remain in effectively capturing the unique visual style of cartoons, characterized by solid colors, simplified textures, and exaggerated character designs. Research explicitly focused on cartoons has begun to address the detection of violent and inappropriate scenes. Knowledge-based systems using character and object databases have been developed to identify violent content in animated videos [22]. Deep learning classifiers, such as Darknet-19, combined with convolutional Long Short-Term Memory (LSTM), have been applied for temporal feature aggregation in cartoons [23]. Multi-modal classifiers that integrate image and text analysis have achieved high accuracy in detecting inappropriate content [21]. Recent advances include sequential deep learning systems that classify violent scenes into multiple categories [24], and modified Faster Region-based Convolutional Neural Network (R-CNN) models with RegNet backbones and attention modules that outperform previous detectors in identifying blood and violence in animated videos [25]. Hybrid CNN architectures combining InceptionV3 and VGG16 have also demonstrated efficiency in recognizing harmful cartoon content with reduced overfitting [26]. Similarly, for real-time blood and violence detection in animated videos, a modified Faster R-CNN with RegNet backbone, deformable convolutions, and attention modules was proposed in [27]. While these contributions represent essential progress in cartoon-specific violence detection, many studies focus on detection within existing scenes rather than predicting future violent events, which is critical for proactive content moderation. Furthermore, the availability of large, diverse cartoon datasets remains limited.

Fusion techniques that combine multiple classifiers or modalities have been employed to enhance detection robustness. Common ensemble strategies such as weighted averaging, majority voting, and stacking help balance the strengths of different models [28]. Temporal prediction methods, including Naive Bayes classifiers, have been used to forecast the likelihood of violence in upcoming scenes based on historical frame patterns, providing early warnings [12]. Similarly, in [29], the authors combine spatial features extracted by EfficientNet with temporal dependencies modeled by Bi-directional Long Short-Term Memory (BiLSTM), integrating both streams to enhance multiclass classification of disturbing YouTube videos by jointly leveraging appearance and motion information. Multimodal rating models that integrate textual, auditory, and visual features offer more reliable violence assessments than unimodal approaches [30]. The reviewed literature demonstrates substantial progress in violence detection for both real-world and animated videos. While models like ResNet18 and VGG16 are widely used for cartoon violence classification, most studies focus on detecting violence within existing scenes rather than predicting future violent events. Moreover, the unique visual characteristics of cartoons and the absence of realistic cues pose challenges that are not fully addressed by models designed for natural videos. Our work addresses these gaps by developing a customized dataset of cartoon violence and employing a fusion of ResNet18 and VGG16 classifiers, combined with a Naive Bayes temporal predictor, to enhance both the detection and prediction of future violent scenes in cartoons, thereby supporting safer content consumption for children.

## 3 Problem Statement

Violence is increasingly common in cartoon programs and can negatively influence children's minds. Unlike real-world videos, cartoons depict violence through characters and scenarios that differ significantly in visual style and presentation. Some cartoon characters are inherently more violent, and scenes often use exaggerated, stylized features that are not present in real footage. These differences render traditional methods for detecting and predicting violence, which are designed for real-world videos, less effective when applied to animated content. Therefore, there is a pressing need for specialized approaches that address the unique nature of cartoons. Numerous machine learning models exist for violence prediction in real-world videos. However, there is a lack of focused research on violence detection in animated movies. The goal of this study is to predict violent scenes in upcoming cartoon segments to enable appropriate filtering, thereby protecting young viewers from harmful content while accounting for the distinctive challenges of animated media.

## 4 Overview of Proposed Framework

This work predicts the presence of violence in upcoming scenes by analyzing patterns of violence in previous scenes and other similar videos. A summary of the model is illustrated in Fig. 1. The model starts by segmenting video clips into shots (i.e., visually similar frames) and scenes (i.e., contextually similar frames) using temporal segmentation. The next step extracts a representative (key) frame from each shot. These key frames are normalized into three layers of RGB images of size 100 × 100. For identification, the key frames are fed into classifiers. This identification process involves utilizing two Convolutional Neural Network (CNN) architectures, ResNet-18 and VGG16, for violence detection. These classifiers are trained using a customized dataset. After classifying the sequence of key frames as violent or non-violent using two distinct convolutional neural network classifiers, the individual outputs are fused using a weighted averaging technique to obtain a consolidated decision. The resulting sequence of classified scenes is then analyzed for patterns of violence, which are utilized to estimate the likelihood of violence in upcoming frames through a probabilistic framework based on the Naive Bayes Classifier. A detailed explanation of each component is provided in the subsequent sections.

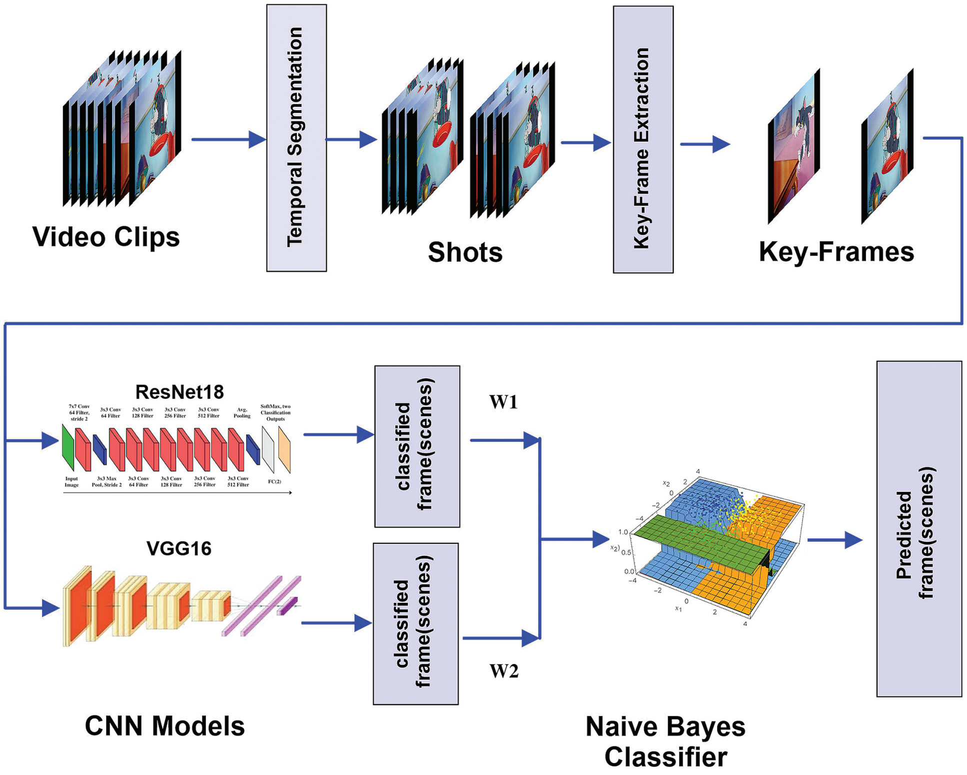


**Figure 1:** Workflow of the proposed framework

### *4.1 Data Processing & Temporal Segmentation*

The proposed model processes RGB-encoded videos by extracting brightness information via a modified integer-to-grayscale conversion, enabling temporal segmentation via a gray-level representation. The non-integer and integer forms of this conversion are given in Eqs. (1) and (2).

$$I_{Gray}(x, y) = 0.299 \cdot I_{Red}(x, y) + 0.587 \cdot I_{Green}(x, y) + 0.114 \cdot I_{Blue}(x, y) \text{ for } x = 1 \ldots \text{Mand } y = 1 \ldots \text{N} \tag{1}$$

$$I_{Gray}(x, y) = \frac{29 \cdot I_{Red}(x, y) + 60 \cdot I_{Green}(x, y) + 11 \cdot I_{Blue}(x, y)}{100} \tag{2}$$

where $I_{Gray}(x, y)$ denotes the gray value at pixel $(x, y)$, and $I_{Red}(x, y)$, $I_{Green}(x, y)$ and $I_{Blue}(x, y)$ represent the corresponding RGB values. The integer form closely approximates the non-integer calculation. This process is illustrated in Fig. 2.

**Figure 2:** RGB to grayscale conversion

Using low-level features such as RGB colors, brightness, and hue, a continuation function is developed to detect shot boundaries within scenes, as summarized in Eqs. (3) and (4).

$$CF = f\left(R, G, B, Gray, Hue\right) \tag{3}$$

$$CF = w_r Red + w_g Green + w_b Blue + w_k Gray + w_h Hue \tag{4}$$

The fixed weights $w_i$ are assigned to different channels. Sub-shots are identified using information from shot boundaries. A key frame is then selected for each sub-shot. Features extracted from these sub-shots are subsequently used to detect and extract objects within the key frames.

### *4.2 Key Frame Extraction & Optimization*

The first step in our framework is to extract representative keyframes from the video to summarize its content effectively. The input video, consisting of $N$ frames, is processed sequentially, with each frame resized to a standard resolution of 240 × 180 pixels to maintain consistency. To detect shot boundaries, we extract low-level visual features from each frame and combine them into a continuous function that reflects changes in the video content over time. Abrupt changes in this function indicate potential shot boundaries, segmenting the video into distinct shots. Fig. 3a illustrates the continuous function over 600 frames of a sample video, while Fig. 3b zooms in on the function for a selected shot, highlighting the points of significant change. From each identified shot, a key frame is chosen to represent the visual content of that segment. The process is summarized in Eqs. (5) and (6).

$$D\left(m\right) = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{F}\left|w_i f_i\left(m\right) - w_i f_i\left(n\right)\right| \tag{5}$$

$$Key\ Frame = k, for\ D\left(k\right) \leq D(\forall)\ for\ all\ \forall \tag{6}$$

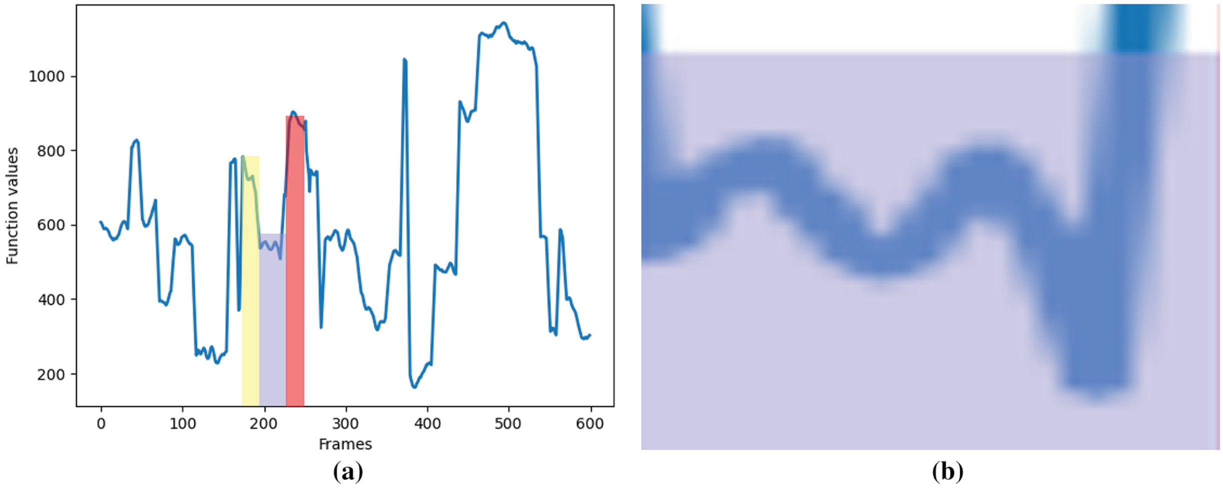


**Figure 3:** Continuous function formation for key frame extraction: (**a**) Continuous function. (**b**) Selected shots

Initial detection often results in numerous false shot boundaries, many of which are closely spaced and do not correspond to meaningful changes in the video. These false boundaries can lead to redundant keyframes, reducing the efficiency and accuracy of subsequent analysis. To address this, we apply an optimization strategy to refine key frame selection. This involves calculating a threshold based on a weighted sum of average feature values across frames, which helps distinguish significant shot boundaries from minor fluctuations. If two consecutive key frames have feature differences below this threshold and are temporally close (fewer than nine frames apart), they are considered redundant. In such cases, the adjacent shots are merged, and a new key frame is extracted to represent the combined segment. This optimization reduces redundancy by eliminating false or insignificant shot boundaries, ensuring that the final set of key frames accurately captures meaningful visual transitions in the video. The process is summarized mathematically in Eqs. (7) and (8). Where $w_i$ is the weight for normalization, $F$ is the total frames in the scene, $N$ is the number of features, $f_i(n)$ is the value of the $i$th feature in frame $n$. By combining feature-based boundary detection with threshold-based optimization, our key frame extraction method strikes a balance between thoroughness and efficiency, providing a concise yet representative summary of the video content for further classification.

$$\tau = \sum_{i=1}^{F} w_i \left( \frac{1}{N} \sum_{r=1}^{N} f_i\left(n\right) \right) \tag{7}$$

$$\sum_{\mathrm{i}=1}^{\mathrm{F}} \mathrm{w_i f_i} \begin{cases} < \tau \text{ ignore the key frame} \\ \geq \tau \text{ accept the frame as a key frame} \end{cases} \tag{8}$$

### 4.3 CNN-Based Classification: ResNet-18 and VGG16 Architectures

To classify scenes as violent or non-violent, we constructed a dataset of 30,415 shots, with a nearly balanced distribution of 50.58% non-violent and 49.42% violent clips. From each shot, a representative key frame was extracted in RGB format and resized to 100 × 100 × 3 pixels. This resulted in 10,521 frames from violent clips and 10,770 frames from non-violent clips. We used 70% of the frames from each class for training and reserved the remaining 30% for testing. Two well-established convolutional neural network (CNN) architectures, ResNet-18 and VGG16, were used for binary classification. These models were selected for their complementary strengths and their suitability for analyzing cartoons and animated content, which often feature distinct visual characteristics such as simplified textures, bold edges, and stylized color patterns. Although more recent architectures may offer improved accuracy and generalization, they typically require greater computational resources and longer training times. By prioritizing VGG16 and ResNet-18, we ensured a practical approach that delivers robust performance without excessive complexity.

#### 4.3.1 Proposed ResNet18 Model

In this study, we propose a ResNet18-based network architecture that accepts input images of size 100 × 75 × 3, as illustrated in Fig. 4. The model begins with an initial convolutional layer that employs filters of size 112 × 112 × 64 and padding [3], followed by subsequent convolutional layers maintaining the exact filter dimensions. Each convolutional layer is succeeded by batch normalization and a rectified linear unit (ReLU) activation function. The first pooling layer employs a window size of 56 × 56 × 64. The final fully connected layer consists of two output nodes for classifying violent and non-violent scenes as binary. This layer includes a softmax node to generate class probabilities and a classification node that computes cross-entropy loss for both standard and weighted objectives, designed explicitly for the two-class problem. This architecture is particularly advantageous for cartoon and animated data because it can efficiently learn complex features while maintaining stable gradient flow, even when the input images contain stylized or exaggerated visual elements. This model strikes a balance between accuracy and computational efficiency, making it well-suited for processing large volumes of animated video frames.

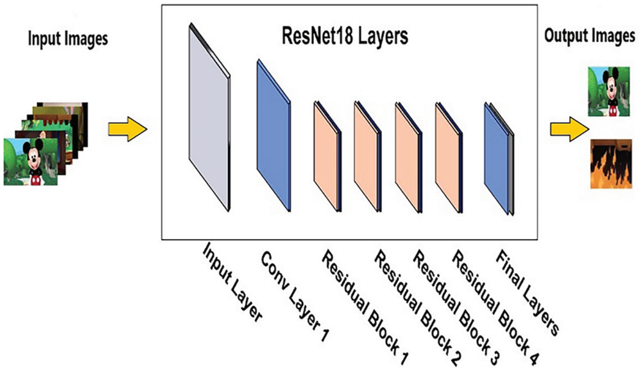


**Figure 4:** Architecture of the proposed ResNet18 model

### 4.3.2 Proposed VGG16 Model

VGG16 features a deeper and more uniform architecture, consisting of 13 convolutional layers with small 3 × 3 filters and three fully connected layers shown in Fig. 5. Its consistent use of small filters allows it to capture fine-grained spatial details and hierarchical features, which is beneficial for cartoons and animations where subtle variations in line work, shading, and color gradients carry essential semantic information. The depth of VGG16 enables it to model these intricate patterns effectively, despite the simplified textures typical of animated content. ReLU activations after each convolutional layer introduce non-linearity, facilitating the learning of complex representations. Although computationally more intensive than ResNet-18, VGG16's ability to extract detailed features complements ResNet-18's efficiency, providing a robust classification framework.

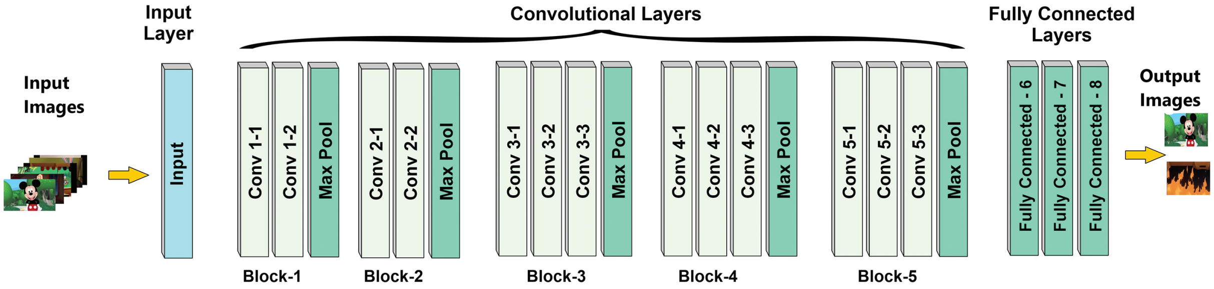


**Figure 5:** Architecture of proposed VGG16 model

## 4.4 Weighted Averaging Fusion Technique

The proposed approach classifies individual key frames as either violent or non-violent by independently training two convolutional neural network models—$ResNet18$ and $VGG16$—on a customized dataset of cartoon imagery. Each model produces a classification for each key frame based on features learned from the customized cartoon dataset. The fusion technique combines these two predictions to improve classification reliability, which can be mathematically formulated as follows. The final prediction probability $P$ for a given key frame is computed as:

$$P = w_1 \cdot P_{ResNet} + w_2 \cdot P_{VGG16} \tag{9}$$

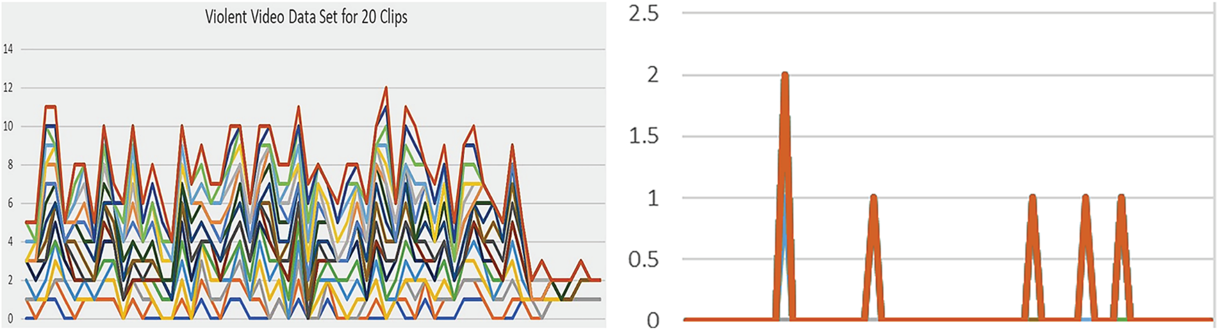
where $P_{ResNet}$ and $P_{VGG16}$ are the predicted probabilities from the respective classifiers, and $w_1$ and $w_2$ are their corresponding weights. These weights are proportional to each model's testing accuracy and satisfy the constraint $w_1 + w_2 = 1$. The fusion process improves decision reliability by addressing cases of agreement and disagreement between classifiers. The process works as follows.

- If both classifiers agree on the predicted label, that label is assigned directly.
- If there is disagreement, the weighted average of their prediction probabilities is calculated using Eq. (9).
- And the label corresponding to the higher weighted probability is then assigned as the final decision.

This method provides a straightforward and computationally efficient way to leverage the complementary strengths of the two models. While more sophisticated ensemble methods (e.g., attention-based fusion or stacking) could yield further gains, weighted averaging ensures interpretability and fast execution, which are essential for our objective of predicting violence in future cartoon scenes in a practical and scalable manner.

### *4.5 Violence Scene Pattern Identification*

To analyze violence patterns effectively, the system first classifies each key frame and its corresponding shot as violent or non-violent. Following this, a sequence of violence intensity levels is constructed to facilitate temporal pattern recognition. During classification, each model outputs a confidence score between 0 and 1, which is discretized into 16 levels, where level 0 indicates the lowest predicted violence intensity and level 16 the highest. Fig. 6a,b illustrates the discretized ranks for 20 violent and non-violent sample clips. To maintain clarity and interpretability, short-duration sequences are used, as presenting full-length results would be overly complex. The resulting violence patterns for both violent and non-violent clips are shown in Fig. 7.



**(a)** **(b)**

**Figure 6:** Violent & non-violent video dataset for 20 clips, (**a**) Violent video dataset, (**b**) Non-violent video dataset

The sequence of violence in each clip is input to the classifier to predict the violence level of the given scene. For a current scene $n$, the neighborhood includes frames from $\boldsymbol{S}_{n-w}$ to $\boldsymbol{S}_{n+f}$. where $w$ and $f$ denote the number of previous and future scenes, respectively. Scenes from $\boldsymbol{S}_{n-w}$ to $\boldsymbol{S}_{n}$ from the feature set, while scene $\boldsymbol{S}_{n+f}$ serves as the classifier's class label. Fig. 8 illustrates this sequence, with $n$ the current scene and scenes from $n - w$ to $n + f$ representing the temporal context (previous and future scenes); here, $w = 5$ and $f = 1$. These sequences of previous and future scenes are extracted to create a dataset of feature vectors paired with corresponding class labels, as shown in Fig. 8. This approach enables the model to learn temporal dependencies in violence patterns, thereby improving scene-level prediction.

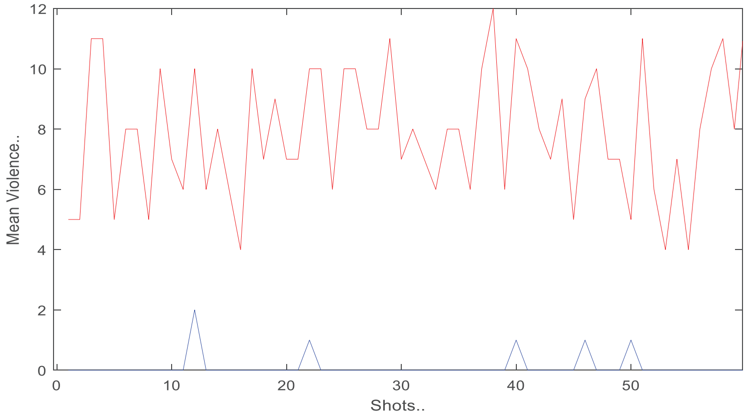
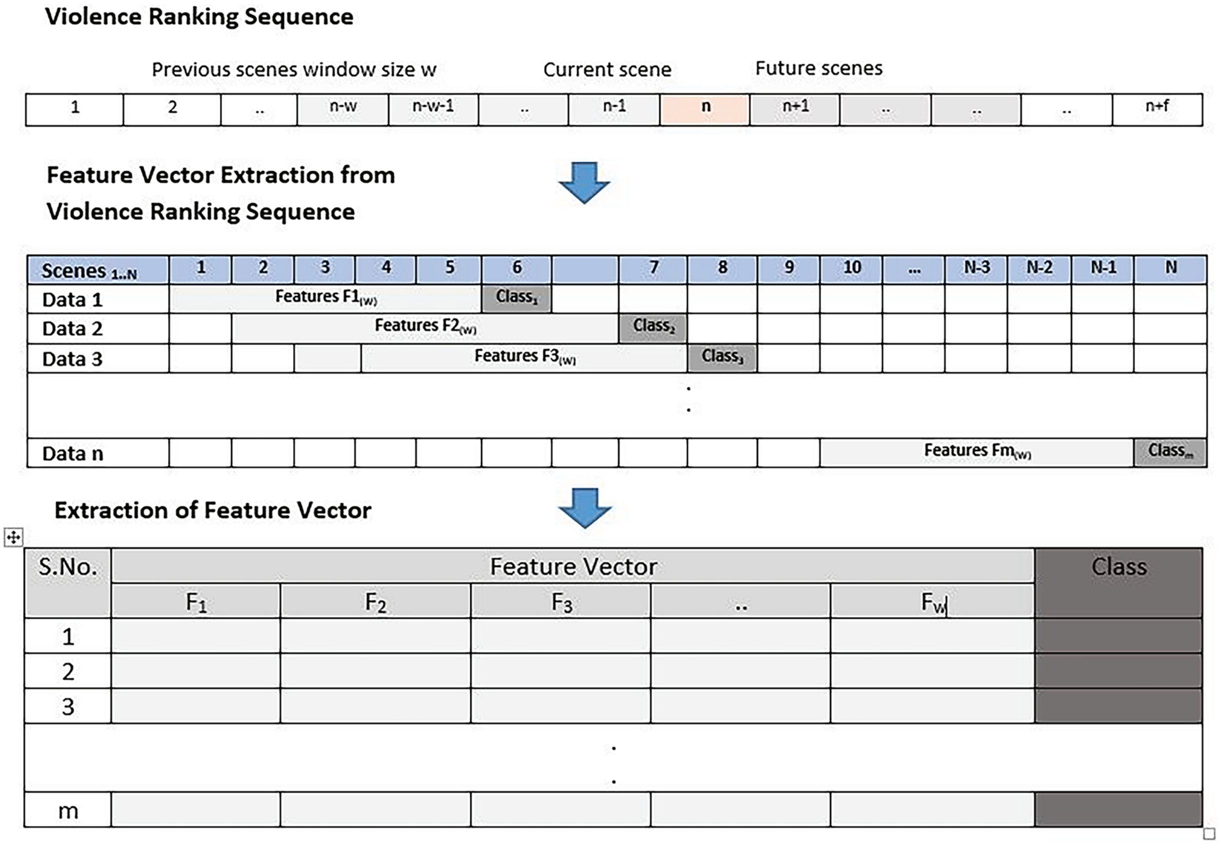


**Figure 7:** Violent & non-violent patterns



Violence Ranking Sequence

| Previous scenes window size w | | | | | | | Current scene | Future scenes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | .. | n-w | n-w-1 | .. | n-1 | n | n+1 | .. | .. | .. | n+f |

Feature Vector Extraction from Violence Ranking Sequence

| Scenes $_{1..N}$ | 1 | 2 | 3 | 4 | 5 | 6 | | 7 | 8 | 9 | 10 | ... | N-3 | N-2 | N-1 | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data 1 | Features F1$_{(w)}$ | | | | | Class$_1$ | | | | | | | | | | |
| Data 2 | | Features F2$_{(w)}$ | | | | | | Class$_2$ | | | | | | | | |
| Data 3 | | | | Features F3$_{(w)}$ | | | | | Class$_3$ | | | | | | | |
| . | | | | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | | | | |
| Data n | | | | | | | | | | | Features Fm$_{(w)}$ | | | | | Class$_m$ |

Extraction of Feature Vector

| S.No. | Feature Vector | | | | | Class |
|---|---|---|---|---|---|---|
| | $F_1$ | $F_2$ | $F_3$ | .. | $F_w$ | |
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| . | | | | | | |
| . | | | | | | |
| m | | | | | | |

**Figure 8:** Process of feature extraction from the violence ranking sequence

### *4.6 Temporal Prediction with Naive Bayes*

In this stage, violence classification leverages temporal context by analyzing sequences of scenes rather than isolated frames. The system constructs a long temporal sequence of violence ranks from multiple videos, concatenated into a single sequence of length 15,059. Using a sliding window approach, feature vectors are formed by extracting the ranks of five consecutive shots. At the same time, the class label corresponds to the violence level of the sixth shot in the sequence, as illustrated in Fig. 8. This dataset enables training a Naive

Bayes classifier to predict the probability of violence in upcoming scenes based on the temporal patterns of preceding frames.

Our choice of the Naive Bayes classifier is motivated by the need for a simple, fast, and computationally efficient method suitable for real-time decision-making based on previous scene history. Unlike more complex models such as LSTMs or Transformers, which require extensive resources and training time to capture intricate temporal dependencies, Naive Bayes offers a lightweight solution ideal for consecutive predictions in cartoon violence analysis. It assumes conditional independence among features, simplifying computation. This approach enables rapid inference without sacrificing the ability to incorporate temporal context, which is essential for real-time cartoon content prediction.

### *4.7 Summarized Algorithm & Computational Complexity*

This section presents a concise summary of the proposed algorithm, along with an analysis of its computational complexity. The complexity assessment in Table 1 focuses on the time and space resources required by each stage, providing insight into the algorithm's efficiency and scalability.

**Table 1:** Complexity of proposed model algorithm

| Steps | Description | Approximate complexity |
|---|---|---|
| 1 | Data processing & temporal segmentation | O(N × h × w × 3) ≈ 0.1G × O(N) (fixed size 240 × 180) |
| 2 | Key frame extraction & optimization | O(N × h × w) ≈ 43,200 × O(N) (grayscale images) |
| 3 | Data augmentation | O(5 × N × 100 × 100 × 3) ≈ 30,000 × O(5N) |
| 4a | ResNet-18 training & testing | ≈ 6.1G × O(N) FLOPs (per image) |
| 4b | VGG16 training & testing | ≈ 46.5G × O(N) FLOPs (per image) |
| 4c | Weighted averaging fusion technique | O(N) (weighted averaging) |
| 5 | Violence scene pattern identification with Naive Bayes classifier | O(n × d × k) ≈ 5 × O(N) (lightweight) |

```
Step 1: Data Processing & Temporal Segmentation
1.  For each frame, $F_i$ do
    a.  $F_i$ (Gray) ← $F_i$ (RGB) // Convert each Frame from RGB to Gray-level
    b.  $I_i \leftarrow \frac{1}{N}\sum_{\substack{x=1\\ y=1}}^{M,N} P(x,y)$ // Calculate Average Intensity
2.  Sig ← $f(I_i)$ // convert average intensity into a continuous signal
3.  $D_{Sig} \leftarrow \frac{d}{dt}Sig$ // calculate the first-order derivative
Step 2: Key Frame Extraction and Optimization
4.  P(i) ← $\wp(|D_{Sig}|)$ // Extract peaks from first-order derivative
5.  OP(j) ← $\nexists(P(i))$ // Remove False peaks (too close, insignificant difference)
6.  DS ← Extract Frames at peaks OP(j) // pick the frames at peaks and develop a dataset
Step 3: Data Augmentation
7.  PDS ← Preprocessing (DS) // Apply preprocessing of the dataset
8.  ADS ← Data augmentation (PDS) // Application of Geometric transformation to augment the data set
```

```
Step 4: Fusion of CNN-Based Classification: ResNet-18 and VGG16 Architectures
Step 4a: Classification Using ResNet18
Net18 ← Training (ResNet18, $ADS_{70\%}$) // Train RestNet18 on 70% of dataset
R1 ← Test (Net18, $ADS_{30\%}$) // Test the trained model using 30% dataset
Step 4b: Classification Using VGG16
Net16 ← Training (VGG16, $ADS_{70\%}$) // Train VGG16 on 70% of dataset
R2 ← Test (Net16, $ADS_{30\%}$) // Test the trained model using 30% dataset
Step 4c: Weighted Averaging Fusion Technique
IF $R1 = R2$ // When both the classifiers agree
    $SR_{sc} \leftarrow R1$ // Result is taken as it is
ELSE // When Both the classifiers disagree
    $SR_{sc} \leftarrow w_1 \cdot R1 + w_2 \cdot R2$ // Fusion of the two classifiers is obtained using a weighted average
ENDIF
Step 5: Violence Scene Pattern Identification with Naive Bayes Classifier
$SRDS(c, f_1 \ldots f_w) \leftarrow SR_{sc} (c \ldots c + w - 1)$ // using windows size w (w = 5) develops scene rank dataset
$PSR(i) \leftarrow$ Predict (NB-Classifier, $SRDS(i - w \ldots i - 1)$) // Using Naive Bayes Classifier predict the next scene based on the previous w scenes (for instance, five previous scenes are used for prediction.
```

$N$ denotes the number of frames or images processed. The image height ($h$) and width ($w$) are fixed at 240 and 180 pixels, respectively, for Steps 1 and 2. Floating-point operations (FLOPs) represent the computational cost of convolutional neural network (CNN) training and testing. Data augmentation increases the dataset size by a factor of five through geometric transformations. The computational complexity of the Naive Bayes prediction depends on the number of samples $n$, the feature dimension $d$, and the number of classes $k$.

## 5 Results and Discussion

This work predicts violence in upcoming scenes by analyzing patterns from previous scenes and similar videos, as summarized in Fig. 1. The process begins with temporal segmentation, dividing video clips into shots (visually similar frames) and scenes (contextually similar frames). From each shot, a representative key frame is extracted and resized to 100 × 100 RGB images. These key frames are then classified as violent or non-violent using two convolutional neural networks, ResNet-18 and VGG16, trained on a customized dataset. The outputs of these classifiers are combined through weighted averaging to produce a consolidated violence prediction for each scene. Finally, the sequence of classified scenes is analyzed using a Naive Bayes classifier to estimate the likelihood of violence in future frames. Detailed descriptions of each component follow in the subsequent sections.

### *5.1 Dataset*

This work introduces a customized cartoon violence dataset created through manual annotation and automated segmentation. Videos were first labeled as violent or non-violent by human reviewers, and then segmented into shots using an algorithm that detects visual changes. Key frames were extracted and optimized to remove redundancy. The dataset comprises 30,415 video shots, divided into 15,030 violent and 15,385 non-violent clips. A 70:30 train-test split was applied, ensuring balanced representation across both sets. The overall distribution is presented in Table 2. Unlike existing real-world violence datasets, this collection focuses on animated content, addressing unique challenges such as stylized characters and context-dependent cues, and establishes a new benchmark for predicting violence in future cartoon scenes.

**Table 2:** Training & testing data

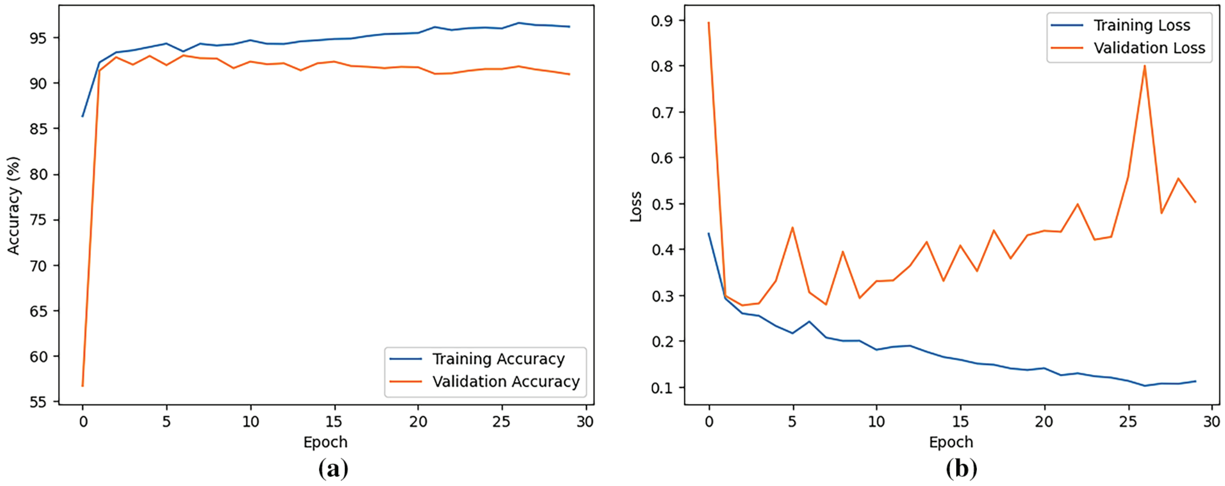
| Class | Total frames | Training (70%) | Testing (30%) |
|---|---|---|---|
| **Violent** | 15,030 | 10,521 | 4509 |
| **Non-violent** | 15,385 | 10,770 | 4615 |

### *5.2 System Configuration*

For training and testing the proposed model, pseudocode was implemented in Python 3 and executed on an Nvidia GTX Titan X GPU with 12 GB of memory. The system also included a sixth-generation Intel Core i5 processor, which provides reliable performance for data processing and general tasks. Storage was handled by a 256-GB SSD, offering faster read/write speeds compared to traditional HDDs, and 8 GB of RAM supported efficient multitasking. The environment was set up with standard Python libraries commonly used in machine learning workflows, including Numpy and Scipy for numerical computations, Pandas and Matplotlib for data analysis and visualization, and TensorFlow with Keras for building and training deep neural networks.

### *5.3 Training Resnet18*

The ResNet18 model was trained for 30 epochs, with the data shuffled at each epoch. Training and validation accuracies reached 96.30% and 92.94%, respectively, after 153 iterations. The training accuracy improved steadily from 86.9% to 96.30% while the training loss decreased from 45.62 to 10.73, as shown in Fig. 9. Validation accuracy was monitored after each epoch to assess generalization. Testing on the remaining 30% of frames yielded an accuracy of 92.49%.



**(a)** **(b)**

**Figure 9:** Performance comparison of the proposed Resnet18 model using accuracy & loss, (**a**) training validation accuracy, (**b**) training & validation loss

### *5.4 Testing Resnet18*

The ResNet-18 model was evaluated using the confusion matrix shown in Fig. 10. The matrix reports 4389 true positives (TP), correctly identified violent scenes, and 4050 true negatives (TN), correctly identified non-violent scenes. False positives (FP) numbered 329, representing non-violent scenes misclassified as violent, while false negatives (FN) totaled 356, indicating missed violent scenes. These values underpin the model's overall accuracy of 92.49%, demonstrating strong predictive performance on the test set. The Receiver Operating Characteristic (ROC) curve in Fig. 10 further illustrates the model's discrimination ability, with an area under the curve (AUC) of 0.9367, highlighting its robust capacity to distinguish between violent and non-violent content. This performance aligns with ResNet-18's reputation as a lightweight yet powerful architecture for image classification, benefiting from residual connections to optimize training efficiency and accuracy.

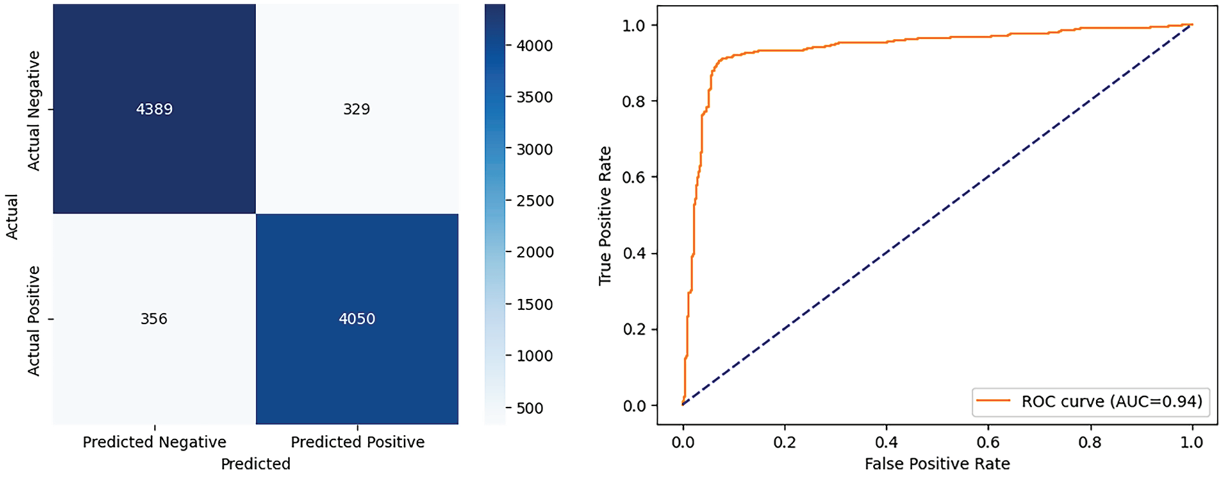


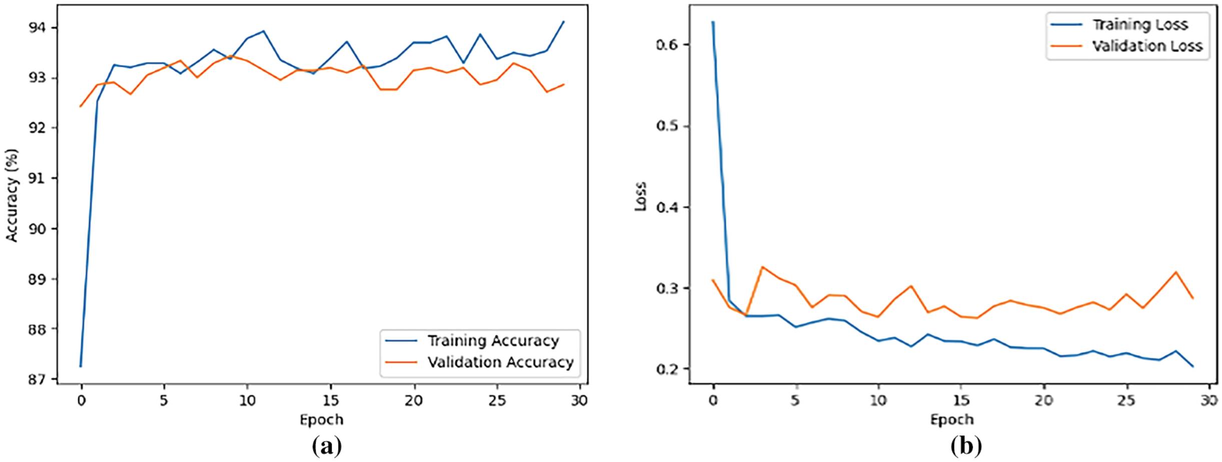
**Figure 10:** Confusion matrix and ROC curve of the proposed ResNet18 model

### *5.5 Training VGG16*

The training accuracy of the VGG16 model over 153 iterations and 30 epochs is shown in Fig. 11a, illustrating its learning progression. The model achieved final training and validation accuracies of 93.42% and 93.66%, respectively, demonstrating strong generalization to unseen data. Validation was performed every 30 iterations. With a learning rate set at 0.01, the model effectively adjusted its parameters during optimization. The accuracy curve shows a steady upward trend, reflecting continuous improvement in classification performance. Meanwhile, the training loss decreased from 68.31 to 24.24, as shown in Fig. 11b, confirming successful optimization. On the held-out test set, the model attained an accuracy of 92.95%.

### *5.6 Testing VGG16*

The classification performance of the VGG16 model during testing is illustrated by the confusion matrix presented in Fig. 12. It correctly identified 4292 true positives (TP) and 4189 true negatives (TN), while producing 286 false positives (FP) and 357 false negatives (FN). These values contribute to an overall accuracy of 92.95%, reflecting the model's strong predictive capability. The ROC curve in Fig. 12 further supports this, with an area under the curve (AUC) of 0.9189, indicating reliable discrimination between classes.



(a)

(b)

**Figure 11:** Performance comparison of the proposed VGG16 model using accuracy & loss (**a**) training & validation accuracy (**b**) training & validation loss

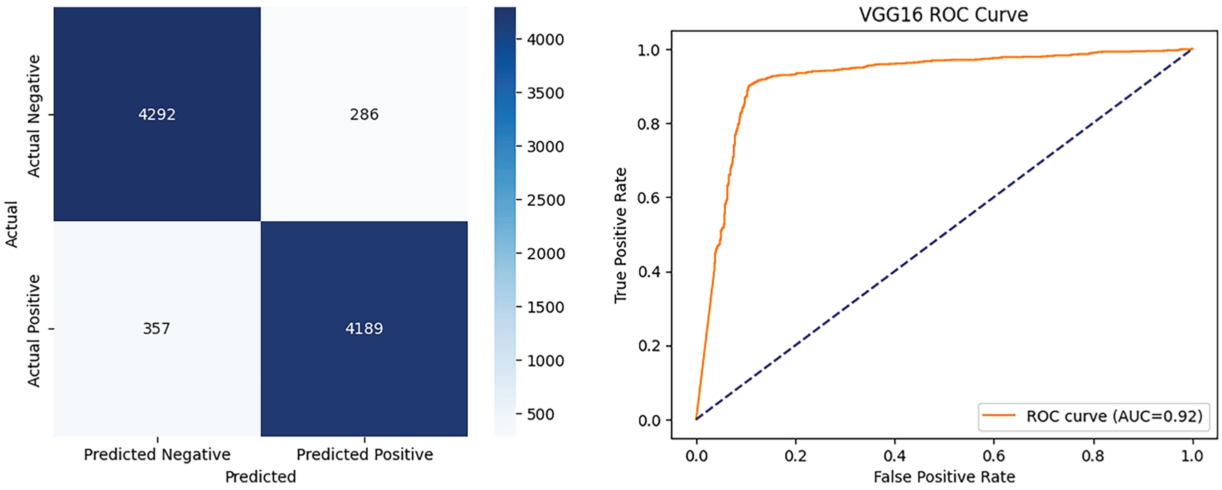


**Figure 12:** Confusion matrix and ROC curve of the proposed VGG16 model

### *5.7 Consolidated & Predictive Accuracy Using Weighted Fusion*

The proposed framework employs a weighted averaging fusion strategy to combine predictions from ResNet-18 and VGG16, assigning weights based on their individual accuracies to achieve a consolidated accuracy of 92.69%, unlike majority voting, which treats models equally and overlooks performance variations—weighted averaging leverages the distinct strengths of each model while remaining simpler and more computationally efficient than stacking. This balance of accuracy, speed, and resource use is particularly suited to cartoon imagery, which features stylized, exaggerated visual elements. Building on this fusion, the aggregated outputs are further processed by a Naive Bayes classifier to predict violence in future scenes. Using a 70:30 training-testing split, the system achieves a predictive accuracy of 92.84%. To capture temporal dynamics, a sliding window of the five most recent scenes is employed, offering a practical trade-off between computational efficiency and reliable detection of both short and extended violent sequences. Larger windows could capture long-term dependencies but increase complexity, while smaller ones risk overlooking sustained patterns. The chosen approach effectively balances these trade-offs, enabling accurate and efficient prediction of violent events in animated content. Overall, integrating ResNet-18 and VGG-16 via weighted fusion, coupled with temporal modeling via Naive Bayes, provides a robust, computationally viable solution for predicting future scene violence in cartoons.

### *5.8 Results and Error Analysis*

This section evaluates the performance of the proposed multi-CNN fusion framework and the Naive Bayes prediction stage, highlighting their effectiveness in detecting and forecasting violent content in animated videos.

Table 3 summarizes the comparative performance of the individual CNN models (ResNet18 and VGG16), their weighted fusion, and the Naive Bayes classifier applied for violence detection in animated videos. All models demonstrate strong classification performance, with accuracies ranging from 92.49% (ResNet18) to 92.95% (VGG16). Precision and recall metrics are similarly high, indicating that the models effectively identify violent scenes while minimizing false alarms. VGG16 achieves the highest accuracy (92.95%) and precision (93.76%), suggesting it is particularly effective at identifying violent content with few false positives. ResNet18 closely follows, with balanced recall (92.50%) and precision (93.09%), indicating its ability to detect violent scenes reliably. The weighted fusion of these two models achieves high accuracy (92.69%) and balances precision and recall, leveraging their complementary strengths to improve overall stability. The Naive Bayes classifier, which uses fused CNN outputs for temporal prediction, achieves an accuracy of 92.84%, with precision and recall values consistent with those of the fusion stage.

**Table 3:** Comparative performance of the Multi-CNN fusion framework & Naive Bayes classifier

| Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| ResNet18 | 0.9249 | 0.9309 | 0.9250 | 0.9279 | 0.9367 |
| VGG16 | 0.9295 | 0.9376 | 0.9232 | 0.9303 | 0.9189 |
| Weighted fusion | 0.9269 | 0.9331 | 0.9205 | 0.9267 | 0.9275 |
| Naive Bayes | 0.9284 | 0.9348 | 0.9244 | 0.9295 | 0.9312 |

This indicates that the temporal modeling effectively forecasts violence in future scenes based on prior fused predictions. Error analysis reveals that false positives often occur in playful or exaggerated fight scenes every day in animated media, such as cartoon characters engaging in competitive or comedic skirmishes. These scenes share visual cues with violent content, leading to occasional misclassification. Conversely, false negatives tend to arise in subtle or low-motion violent acts, such as threatening gestures or the presence of a weapon without overt action, which are harder for the models to detect. Overall, the high precision across models ensures that non-violent scenes are rarely misclassified as violent, reducing false alarms. Meanwhile, strong recall values confirm that most violent content is successfully identified. The fusion approach and Naive Bayes prediction stage together enhance robustness and temporal forecasting, making the framework well-suited for practical violence detection in animated videos.

### *5.9 Performance Comparison with Existing Models*

To evaluate the effectiveness of the proposed framework, we compared its performance with that of existing studies on violence detection in cartoons, animated videos, and general videos. Prior research has employed diverse strategies, including multimodal deep learning integrating metadata and visual features [21], ConvLSTM-based temporal modeling [23], lightweight CNN–LSTM hybrids [24], and transformer-based temporal architectures [20]. Other notable approaches involve spatiotemporal feature extraction using 3D CNNs with Inception-ResNet [15] and video transformer networks combined with ConvLSTMs on mixed datasets [27].

As summarized in Table 4, reported accuracies vary widely across these studies. Earlier multimodal methods, such as Chuttur and Nazurally [21], achieved moderate accuracy around 76%. More recent temporal models based on ConvLSTM and hybrid architectures [23,24], as well as transformer-based methods [20], have reported accuracies exceeding 95%. Singh and Tyagi [28] reported the highest accuracies, with VGG16 achieving 99.98% on the Hockey Fight dataset and 99.94% on the Real-Life Violence Situations dataset. Similarly, Chandira Prabha et al. [16] achieved 96% accuracy on the Kaggle Real-Life Violence Dataset. In comparison, Shen and Zou [15] demonstrated strong spatiotemporal performance with 94.3% and 95.1% accuracy on the UCF-101 and HockeyFights datasets, respectively.

**Table 4:** Performance comparison with existing models

| Reference | Method | Dataset | Accuracy (%) |
|---|---|---|---|
| Chuttur & Nazurally (2022) [21] | Multi-modal deep learning (CNN + metadata) | Proprietary cartoon dataset | 76.00 |
| Fadzli et al. (2022) [23] | ConvLSTM (Darknet-19 + Conv-LSTM) | 300 cartoon clips | 96.43 |
| Khan et al. (2024) [24] | Lightweight CNN with LSTM | Cartoon violent key frames (~12k, 5 classes) | 97% |
| Ahmed et al. (2023) [27] | VTN/I3D/ConvLSTM | MOB dataset (cartoon + gameplay) | 77.9/72.1/69.7 |
| Shen & Zou (2020) [15] | 3D CNN with Inception-ResNet | UCF-101, HockeyFights | 94.3/95.1 |
| Chandira Prabha et al. (2025) [16] | Hybrid (MobileNetV2 + ConvLSTM) | Kaggle real-life violence dataset (1000 clips, 50/50 violent vs non-violent) | 96.0 |
| Chatterjee et al. (2025) [20] | Temporal-aware Transformer (MobileTransformerSeq) | YouTube violence/non-violence dataset (2000 videos) | 97.20 |
| Singh & Tyagi (2023) [28] | CNNs (VGG-16, VGG-19, InceptionV3, MobileNetV3 Small) | Hockey fight dataset | 99.98 (VGG-16 best) |
| Singh & Tyagi (2023) [28] | CNNs (VGG-16, VGG-19, InceptionV3, MobileNetV3 Small) | Real life violence situations dataset | 99.94 (VGG-16 best) |
| Proposed framework | Weighted Fusion (ResNet-18 + VGG16) + Naive Bayes | Cartoon video dataset (70:30 split) | 92.84 |

By comparison, our proposed weighted fusion of ResNet-18 and VGG-16 with Naive Bayes achieved 92.84% accuracy on a proprietary dataset of cartoon violence. While slightly lower than the top-performing benchmarks on real-world datasets, this result highlights the framework's robustness within the relatively underexplored domain of animated and cartoon violence detection. More critically, our approach advances beyond classification by addressing the prediction of future violence in animated content, a novel contribution not explored in prior studies. The fusion design leverages the complementary strengths of ResNet-18 and VGG-16, striking a balance between accuracy, scalability, and computational efficiency.

## 6 Conclusion

This paper presents a novel multi-CNN fusion framework for detecting and predicting violent content in animated media, addressing the critical challenge of safeguarding children from inappropriate digital content. By integrating ResNet-18 and VGG16 via a weighted average fusion technique, the framework improves scene classification accuracy. The fused outputs are subsequently analyzed using a Naive Bayes classifier over a sliding window of prior scenes, enabling effective prediction of future violent content. Experimental results on a custom dataset of 30,415 annotated shots demonstrate a predictive accuracy of 92.84%, highlighting the framework's robustness and applicability for real-world content moderation systems. The proposed approach balances accuracy, computational efficiency, and interpretability, making it well-suited for deployment in resource-constrained environments typical of animated video analysis.

Further, Naive Bayes serves as an effective baseline for temporal modeling. Therefore, future research will investigate advanced sequential models such as Long Short-Term Memory (LSTM) networks and Transformers, to capture more complex temporal dependencies. Additionally, expanding the dataset and incorporating richer scene descriptors—including morphological and motion features—is anticipated to further improve model generalization and prediction performance. In summary, the multi-CNN fusion framework establishes a strong foundation for predictive violence detection in animated media. It offers promising directions for future enhancements in both model sophistication and dataset diversity.

**Acknowledgement:** We acknowledge the support from Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R138), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**Funding Statement:** This study was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R138), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**Author Contributions:** The authors confirm their contribution to the paper as follows: study conception and design: Tahira Khalil; data collection: Muhammad Imran Khan Khalil; analysis and interpretation of results: Tahira Khalil, Sadeeq Jan; draft manuscript preparation: Tahira Khalil; manuscript revision: Rania M. Ghoniem, Muhammad Imran Khan Khalil; overall supervision: Sadeeq Jan; funding acquisition: Rania M. Ghoniem. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The datasets and materials used or analyzed during the current study are available from the corresponding author on reasonable request. All relevant data supporting the findings of this study have been included in the manuscript.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Paron GC. Igniting learner's behavior and language development through cartoon movies: a phenomenological study. Am J Multidis Res Innov. 2022;1(2):30–3. doi:10.54536/ajmri.v1i2.251.
2. Ghosh S, Panchal D, Shah P. A comparative study of overt aggression & moral judgement in relation to age among cartoon watchers & non-cartoon watcher students. Gap Bodhi Taru A Glob J Humanit. 2020;3(2):1–12.
3. Susanu N. Violence, the effect of cartoons on the child's emotional development. New Trends Psychol. 2022;4(1):114–9.

4. Ghilzai SA, Alam R, Ahmad Z, Shaukat A, Noor SS. Impact of cartoon programs on children's language and behavior. Insights Lang Soc Cult. 2017;2(1):104–26.
5. Krcmar M, Cingel DP, Vigil SL, Snyder AL. The effects of fantastical characters and prosocial message duration on children's narrative, emotional, and moral lesson comprehension of a cartoon. Mass Commun Soc. 2024:1–26. doi:10.1080/15205436.2024.2404626.
6. Yan W, Mobai C, Tong W, XiaoBo Y. The negative influence of violence graphic animations cartoon to children under age of 12: a review. J Leg Ethical Regul Issues. 2023;26:1.
7. Abd elbaseer Mahmoud D, Mahmoud Zaki M, Ahmed Mostafa H. Correlation between cartoon violence, aggressive behavior and psychological well-being among primary school children. Egypt J Health Care. 2022;13(1):1232–51. doi:10.21608/ejhc.2022.225113.
8. Parvin F, Islam S. The impact of cartoon programs on children's physical health, intelligence, behavior and activities. Eur J Physiother Rehabil Stud. 2020;1(1):1–22.
9. Khan SU, Haq IU, Rho S, Baik SW, Lee MY. Cover the violence: a novel deep-learning-based approach towards violence-detection in movies. Appl Sci. 2019;9(22):4963. doi:10.3390/app9224963.
10. Martinez VR, Somandepalli K, Singla K, Ramakrishna A, Uhls YT, Narayanan S. Violence rating prediction from movie scripts. Proc AAAI Conf Artif Intell. 2019;33(1):671–8. doi:10.1609/aaai.v33i01.3301671.
11. Jaiswal SG, Mohod SW. Recapitulating the violence detection systems. In: Proceedings of the ICCCE 2019; 2019 Feb 1–2; Pune, India. Singapore: Springer Singapore; 2019. p. 197–203. doi:10.1007/978-981-13-8715-9_25.
12. Traore A, Akhloufi MA. Violence detection in videos using deep recurrent and convolutional neural networks. In: Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2020 Oct 11–14; Toronto, ON, Canada. doi:10.1109/smc42975.2020.9282971.
13. Karkehabadi A, Oftadeh P, Shafaie D, Hassanpour J. On the connection between saliency guided training and robustness in image classification. In: Proceedings of the 2024 12th International Conference on Intelligent Control and Information Processing; 2024 Mar 8–10; Nanjing, China. Piscataway, NJ, USA: IEEE; 2024. p. 1–6. doi:10.1109/ICICIP60808.2024.10477811.
14. Taha MM, Alsammak AW, Zaky AB. Filtering of inappropriate video content: a survey. Int J Eng Res Technol. 2022;11(2):325–32.
15. Shen J, Zou W. Violence detection based on three-dimensional convolutional neural network with inception-ResNet. In: Proceedings of the 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS); 2020 Dec 11–13; Shenyang, China. doi:10.1109/tocs50858.2020.9339755.
16. Chandira Prabha S, Kaviyadharshini S, Micheal AA. A hybrid deep learning model for violence detection. In: Proceedings of the 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT); 2025 Feb 5–7; Bengaluru, India. doi:10.1109/IDCIOT64235.2025.10914952.
17. Pei Y, Wang Z, Chen H, Huang B, Tu W. Video scene detection based on link prediction using graph convolution network. In: Proceedings of the 2nd ACM International Conference on Multimedia in Asia; 2021 Mar 7; Virtual. doi:10.1145/3444685.3446293.
18. Jain N, Gupta V, Shubham S, Madan A, Chaudhary A, Santosh KC. Understanding cartoon emotion using integrated deep neural network on large dataset. Neural Comput Appl. 2022;34(24):21481–501. doi:10.1007/s00521-021-06003-9.
19. Aldahoul N, Karim HA, Abdullah MHL, Wazir ASB, Ahmad Fauzi MF, Tan MJT, et al. An evaluation of traditional and CNN-based feature descriptors for cartoon pornography detection. IEEE Access. 2021;9:39910–25. doi:10.1109/access.2021.3064392.
20. Chatterjee R, Roy Choudhury R, Kumar Gourisaria M, Banerjee S, Dey S, Sahni M, et al. Temporal-aware transformer approach for violence activity recognition. IEEE Access. 2025;13:70779–90. doi:10.1109/access.2025.3560828.
21. Chuttur MY, Nazurally A. A multi-modal approach to detect inappropriate cartoon video contents using deep learning networks. Multimed Tools Appl. 2022;81(12):16881–900. doi:10.1007/s11042-022-12709-2.
22. Khalil T, Bangash JI, Khan AW, Lashari SA, Khan A, Ramli DA. Detection of violence in cartoon videos using visual features. Procedia Comput Sci. 2021;192(1):4962–71. doi:10.1016/j.procs.2021.09.274.

23. Fadzli MAHM, Hassan MFA, Ibrahim N. Explicit kissing scene detection in cartoon using convolutional long short-term memory. Bull Electr Eng Inform. 2022;11(1):213–20. doi:10.11591/eei.v11i1.3542.
24. Khan NF, Amin SU, Jan Z, Yan C. Detection of violent scenes in cartoon movies using a deep learning approach. IEEE Access. 2024;12(1):154080–91. doi:10.1109/ACCESS.2024.3480205.
25. Tang Y, Chen Y, Sharifuzzaman SASM, Li T. An automatic fine-grained violence detection system for animation based on modified faster R-CNN. Expert Syst Appl. 2024;237(5):121691. doi:10.1016/j.eswa.2023.121691.
26. Mallidi SKR, Bellapukonda S, Gollapudi N, Malaka K, Sreeshanth P, Polisetti SSH. Deep learning in cartoon moderation: distinguishing child-friendly content with CNN architectures. In: Proceedings of the Cognitive Computing and Cyber Physical Systems; 2024 Apr 5–7; Bhimavaram, India. Cham, Switzerland: Springer Nature Switzerland; 2025. p. 241–56. doi:10.1007/978-3-031-77075-3_20.
27. Ahmed SH, Khan MJ, Qaisar HMU, Sukthankar G. Malicious or benign? Towards effective content moderation for children's videos. arXiv:2305.15551. 2023.
28. Singh S, Tyagi B. Computational comparison of CNN based methods for violence detection; 2023. [cited 2025 Jan 1]. Available from: https://www.researchsquare.com/article/rs-3130914/latest.
29. Yousaf K, Nawaz T, Habib A. Using two-stream EfficientNet-BiLSTM network for multiclass classification of disturbing YouTube videos. Multimed Tools Appl. 2024;83(12):36519–46. doi:10.1007/s11042-023-15774-3.
30. Xiang T, Pan H, Nan Z. Video violence rating: a large-scale public database and a multimodal rating model. IEEE Trans Multimed. 2024;26(112):8557–68. doi:10.1109/TMM.2024.3379893.