OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 02 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

DEEP CONVOLUTIONAL NETWORK FOR AUTOMATIC VIOLENCE DETECTION IN SURVEILLANCE VIDEOS USING TRANSFER LEARNING

Muhammad Qasim Khan
Department of Computer Science Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan.
Sohail Nawaz Sabir
Business Applications & Database Manager – Middle East, Veolia Water Technologies Ltd, Saudi Arabia.
Fazal Malik*
Department of Computer Science Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan.
Muhsin Khan
Department of Computer Science Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan.

****Corresponding author: Fazal Malik (fazal.malik@inu.edu.pk)***

Article Info

Abstract

Detections of violence within surveillance footage enable instant intervention support that advances public security measures. Traditional systems experience various implementation problems which include redundant frames as well as dependency on datasets and suboptimal generalization capacity. This paper presents an automatic violence detection system which applies Inception-v3 with transfer learning features to address existing system limitations. The system successfully performs effective and precise classification of video frames between violent and non-violent categories. The approach is a five-phase framework consisting of: (1) dataset collection (Hockey Fight and Real-Life Violence datasets), (2) elimination of redundant frames employing unsupervised learning and Euclidean distance-based similarity measures, (3) splitting of datasets (70% training, 30% testing), (4) transfer learning with Inception-v3, fine-tuned by substituting its last three layers, and (5) performance assessment. The system utilizes CNN-based feature extraction and the Stochastic Gradient Descent with Momentum (SGDM) optimizer for training. The proposed system delivers a performance rate of 73.86% and 78.61% on Hockey Fight and Real-Life Violence datasets surpassing other methods in state-of-the-art research. The proposed system improves processing speed and feature learning stability through redundant frame reduction techniques. The system demonstrates strong versatility among different situations which makes it suitable for use in real surveillance applications and content moderation systems. The proposed approach tackles problems with present approaches by developing a scalable solution for effective violence detection. The authors plan to focus on hyperparameter optimization and dataset testing as future work for accuracy enhancement.

Keywords:
Surveillance, redundant frame removal, Inception-v3, transfer learning, deep learning, violence detection.

1. Introduction

The recognition of violent behaviors remains vital for understanding movements between humans within security videos because of its essential role in crime prevention. Into smart cities it brings public security improvements combined with crime reduction capabilities. Real-world surveillance systems need both speed and precision of violence recognition systems to permit intervention as soon as possible. The monitoring process through closed-circuit television (CCTV) normally occurs after incidents to support legal investigations although it does not stop wrongdoing from taking place as it happens. The monitoring of lengthy surveillance footage campaign depends heavily on minimal personnel but suffers from the unreliability of human visual oversight because of prolonged shifts. Video processing through deep learning produces instant analysis which assists patient care and prisoner oversight and public defense alongside security activities. Complex machine learning systems encounter difficulties with big datasets combined with multiple classes which reduces their current best performance levels [1]. Organizations and public entities alongside law enforcement agencies use broad-reaching observance methods to both detect threats and handle violent altercations because of increased criminal conduct. Detecting violence automatically remains a complex problem that affects successful detection. A Convolutional Neural Network operates proficiently to identify multiple thousands of actions at once but its effective development needs expert involvement. The detection system used for violence identification operates in multiple settings which include indoor and outdoor areas through building observability and traffic control activities and police camera systems as shown in Figure 1 [2].

Figure 1. Real-time violence detection in: A) Indoor surveillance, B) Traffic monitoring, C) Police body cameras [2].

1.1. Applications of Violence Detection Systems

The processing of online material constitutes a major aspect of Closed-Circuit Television (CCTV) image management. Media data filtering has become essential for digital times because it enables three critical functions: maintaining control for parents, providing content ratings and preventing unwanted media distribution. The cloud serves as a storage facility for all video audio and animation data which manage large-sized files. Professional experts mark data records by hand so machine learning systems can identify irregularities through their automated process.

1.2. Challenges in Action Recognition

In machine learning the area of action recognition maintains three distinct categories [3]: which include 1) Action Recognition in the Wild – This detects motions while accounting for each background element

while performing recognition. 2) Skeleton Base Recognition – The system employs depth images together with disparity information and human skeletons retrieved from Microsoft Kinect to conduct action classifications. 3) Human Segmentation – divides human bodies detected in video frames before sending them to recognition models for action classification and labeling.

The analysis of videos through automation enables better system mechanisms for content filtering and recommendations. At the same time it optimizes content ranking.

Real-time special-purpose computer systems function as the basis for embedded sensor-based human action recognition. The device operates at a speed which hinders its ability to deliver both high accuracy and efficiency in operations. We need to run machine learning model training sessions at a system with Graphics Processing Units (GPU) while using many different classes which contain thousands of video frames. The model requires verification after training before it becomes available for publication. The networked camera operates by snapping images that get verified against the trained model for real-time action recognition [4].

1.3. Factors Affecting VDS Performance

Real-time classification of human activity represents a challenge for action recognition because it occurs during actual events especially in surveillance videos. Several factors including video quality issues and insufficient lighting conditions as well as lack of contextual element for violence versus non-violence distinction create this detection challenge. Audio surveillance systems become more challenging to detect actions due to the existing operational limitations [5].

1.3.1. Redundant Frames in Surveillance Videos

System creation becomes complicated because real-time violence detection faces various system formation factors. The elimination of redundant frames creates problems for systems based on video recognition because most videos operate at 30 frames per second. The chosen video encoding algorithms establish different frame rates but 30 frames per second contain redundant video information. A proposed method to decrease video redundancy involves utilizing the fifth frame method or keeping only the odd or even frames. The methods implemented have shown limited success in the process [6].

1.3.2. Dataset Dependency and Overfitting

The requirement for a forceful detection mechanism is necessary to detect violence because it can emerge at any moment or place using poor visuals. Most VDS research uses hand-made features and traditional detection models that consume significant time for constructing and further result in excessive redundant features. A high level of redundancy in data features leads to accuracy degradation as well as overfitting in classifier training processes [7].

1.3.3. Feature Representation Limitations

Shape together with color along with other basic features exhibit non-robust properties when recognizing action sequences. Traditional features often fail to maintain accuracy over transformations like brightness together with scale and rotation and conversion changes because they are non-invariant. Some classification methods have proven ineffective to perform successful classification.

Our solution tackles these obstacles efficiently according to the information presented in this paper.

1.4. Deep Learning and CNN

Deep learning functions as part of machine learning through neural networks to discover apps that show complex nonlinear data relationships by running computations across multiple layers of processing. Medical sciences [8, 9], and computer networks [10], and student performance prediction [11], together with Google stock prediction [12, 13], and software engineering [14], apply deep learning and machine learning solutions.

1.4.1. CNNs for Video and Image Analysis

CNNs function similarly to the human visual cortex thus they deliver exceptional performance for image and video processing. The supervised deep learning system Recurrent Neural Networks (RNNs) serves various applications in computer vision and bioinformatics together with Natural Language Processing (NLP) while boosting accuracy and machine performance [4].

1.4.2. Feature Extraction in CNNs

The models train themselves to retrieve essential data characteristics autonomously which makes them free from human-made heuristics. A feed-forward structure defines CNNs while their data flow occurs in one direction only. RNNs implement two-way information transfer between forward and backward nodes. Similar to human neurological cells which transmit vast information, CNN nodes serve as artificial brain cells that connect through dendrites. The CNN model builds its layers on top of each other while its mini-data structures flow from one layer to the next to create a forward-directed artificial network. During training the classification layer improves weights through back-propagation in order to minimize errors while this method changes system operations without performing a re-training process [15].

1.4.3. Evolution of CNNs

CNNs delivered highly accurate results from the 1990s for handwriting recognition and face identification technology. The popularity of CNNs grew when AlexNet became the first model to win the ImageNet competition [16]. Relationship testing between CNN models occurs on the ImageNet benchmark dataset which contains 1,000 different classes. Current deep neural models including RNNs together with CNNs demand substantial dataset information for their training process and show poor results with limited datasets.

The way DL operates to detect specific features through its automated learning process makes it more advantageous than traditional ML for medical image processing applications including diagnosis and segmentation tasks. The detection of tumors benefits from CNNs although successful implementation requires substantial data resources and optimal parameter adjustments. The data shortage barrier gets resolved through transfer learning which incorporates pre-trained ImageNet models to deliver better accuracy results. The Inception-v3 and other pre-trained CNN networks can be adjusted by MATLAB through the replacement of their final layers during TL. CNN structure contains the sequence of convolutional, pooling, normalization, ReLU, sigmoid, fully connected and SoftMax layers and classification layers. An optimization procedure uses learning rate selection to achieve optimal training performance [17].

1.5. Foreground Detection

The fundamental tracking technique in computer vision uses foreground detector as its basis. The process begins by selecting a background picture for reference then establishing a threshold constant for

subtraction and removing image elements from this background before identifying the foreground areas. The foreground detector helps reduce redundant frames that occur in high FPS video recordings. A chosen image functions as the foreground while the reference consists of separated foreground images according to the algorithm [18].

1.6. Optical Flow

Optical flow describes how moving objects appear in different videos through patterns of movement. A video file benefits from optical flow which excellence at both boundary and surface detection enables recognition through scene movements. The velocities of moving objects constitute optical flow while the computation of these velocities depends on brightness patterns in images. The movement of objects is measured by tracking video frames together with instantaneous velocity measurements of each image displacement [19].

1.7. Inception-v3 Model

Inception-v3 represents a CNN model which received training through around 100 thousand images from the Image-Net database. Image-Net provides researchers with its benchmark collection of 1000 different categories. The Inception-v3 neural network operates with 316 layers in two major parts. Layers from 1 to 313 in the model operate through convolution and the following three layers 314-316 perform classification by utilizing fully connected then soft max before generating classification results [6].

1.8. Transfer learning

The process of applying transfer learning refers to taking a pre-trained CNN model and removing its initial bottom three layers so you can create a different model from the same CNN using new data conditions in a training environment. The research involves removing layers 314 to 316 from Inception-v3 along with new layers for violent and non-violent scene detection. Computers need to develop patterns which they can apply or operate on images to extract meaning while learning various conceptual segments. The retraining process of an existing model focuses on recognizing different high-level characteristics. The approach leads to greater precision levels together with decreased training requirements. The training duration along with resource requirements make this technique highly valuable [20].

The detection of violence through automatic surveillance video analysis proves complex because of excessive frames and method limitations and dataset requirements. The present techniques struggle to extract features properly while requiring high computation and producing poor results during dynamic processes. The current deep learning models lack robustness along with real-time performance. A deep learning framework needs development for optimal violence detection capabilities along with high computational efficiency as well as broad dataset generalization.

The main objective exists to develop a violence detection system through the implementation of an optimized deep learning model with transfer learning capabilities. The project targets accuracy excellence, computational efficiency through frame skipping, universal model adaptability for datasets together with reduced false alarms and real-time security system implementation.

Five main steps constitute the proposed methodology which begins with data acquisition and preprocessing followed by Hockey Fight and Real-Life Violence dataset use, normalization then augmentation and frame extraction procedures. The system calculates similarities between frames until all redundant frames are removed from the dataset. A dual-path CNN serves to extract spatial and

temporal features during the third step of the methodology. The optimization process involves using SGDM optimizer for Inception-v3 model refinement and transfer learning through replacement of its final three layers. The data split for training must consist of 70% data while the remaining 30% represents testing needs. Running cross-validation during the training process followed by performance optimization. Analysis includes a performance evaluation of the system based on accuracy and precision along with recall and F1-score measurements alongside state-of-the-art methodology monitoring.

Improved accuracy results from implementing a dual-path CNN together with spatial-temporal optimized feature extraction while frame removal reduces processing demands by 30%. The approach offers high-level generalization for various datasets combined with optimized parameters through system tuning and knowledge transfer capabilities and direct security system deployment capabilities.

The subsequent sections follow with research description on violence identification through a Literature Review in Section 2 while Section 3 shows the proposed methodology then items research implementations through Section 4 that integrates evaluation metrics with results before Section 5 presents major findings in Conclusion and future research paths.

2. Literature Review

The state-of-the-art approaches for detecting violence in surveillance videos rely on three key elements including audio feature extraction and Mo SIFT descriptors for spatiotemporal analysis as well as optical flow motion vectors. Current accurate methods provide 85% success rates but struggle with low precision rate and high resource consumption as well as high computational expenses. Current methods fail to meet operational requirements in real surveillance applications because they produce weak precision rates at unacceptable delays in result production [21].

The computer vision system is 95% accurate in detection on particular datasets with real-time performance. The combination of traditional image processing with deep learning techniques yields this detection system. Studies about detecting violence in videos remain relatively rare among other research endeavors in video analysis. An action recognition framework has introduced based on dense trajectories combined with descriptors which uses optical flow motion boundary detection to find motion trajectories [22]. The approach benefits from camera motion correction together with Fisher Vector encoding [15], emerging as state-of-the-art at its time, this approach involves aggregating multiple traditional descriptors into a bag of features, as the state-of-the-art when it first released since it aggregates multiple classic descriptors into a feature bundle.

The classifier establishes the connections between descriptors to represent different action types as the main step towards automatic video feature extraction. The proposed 3D-based CNN system functionalized action recognition through multiple accompanying video frames that delivered motion data to the network processing system. Thus the approach performed equally well as dense trajectories using frames reduced to one-fourth of resolution quality. The action recognition process primarily focuses its analysis on basic features. The standard approach for feature extraction identifies interesting points using gradient-based and optical flow methodologies and intensity-related and additional local elements [20]. Research studies in the past have implemented threshold criterion for measuring audio and visual parameters.

The researchers determined both auditory acoustic signal amplitudes and energy amounts while analyzing the auditory features. Successfully detecting fast activities such as blood detection through visual features requires the whole entropy analysis of quick variations that evaluates dynamic feature activities based on pixel color thresholds [18]. Some researchers developed acoustic detection systems to

identify fundamental audio occurrences including explosions and gunshots and breaking car glass and vehicle engine noise. The implementation utilizes Hidden Markov Models (HMM) to identify target sound events after which Gaussian mixture models generate semantic associations by analyzing event relationships [21].

The previous systems employed unique specific observed events during their operation. Such method attempts both classification and categorization of individuals based on their actions. The researchers selected low-level features including Space-Time Interest Points (STIP) and Motion SIFT (Mo SIFT) instead of using the Bag of Visual Words (BoVW) method for their work. The method included a histogram showing local motion which was appended to the system design. The video features got transformed into a bag of words through SVM which performed the classification process [23]. The classification of video scenes as violent or normal was achieved through a Local Spatiotemporal Features-based Bag of Visual Words method. The STIP-detected descriptors were selected several times to include spatiotemporal features which ultimately resulted in the formation of frame-based feature collections. The classification of videos with linear SVM generated reasonably better results for particular activities. The identification of violence requires an inclusion of spatial and temporal features together with movement analysis according to all investigated methods [24].

Each study generates outcomes using individual datasets that implement different measurement methods. The variations between theories about violence create barriers when attempting direct evaluations with established methods. Human action analysis automatically continues to evolve into an emerging research field despite the substantial work conducted in action recognition. Studies implement both distinctive sounds and general action descriptors when combining them for detecting fast changes in motion patterns. These studies build their approach on BoVW by combining multiple descriptors for visual word representation. The research explores how spatiotemporal data identification of violence works by uniting sound attributes with video movement analysis yet relies on manually developed descriptors for constructing their heavily subjective model. Machine-learning techniques deliver outstanding results for image processing and video classifying functions primarily through CNN implementations [18, 25].

Elevated machine learning technology enables developers to build COVID-19 pneumonia diagnostic systems as well as sentiment analysis systems and accident prediction systems and cybersecurity systems. The research combines GitHub X-ray data augmentation with optimized Random Forest and AdaBoost and XGBoost and Convolutional Neural Networks (CNNs) for COVID-19 detection [26-29]. XGBoost with AdaBoost and Artificial Neural Networks (ANNs) make up three methods that enhance review classification in the Google Play Store platform [30]. The Random Forest model provides superior output than AdaBoost when applied to dark data-driven crash detection duties [31]. The combined system of XGBoost and AdaBoost within cybersecurity procedures improves URL detection through minimized occurrences of incorrect positive and negative outcome classes. A comprehensive threat detection system operates through structured sequences for better cyber threat discovery capabilities [32-34].

These networks surpassed BoVW as well as other previous methods in diverse challenging datasets. The action recognition community explores these networks for their optimization. The different architectures used for fusing spatiotemporal information have distinct methods for transmitting motion data thus producing several features for the classification operation. Violence recognition research in the current period shows positive results from applying convolutional networks in this field. Most approaches extract neural network features followed by added handmade features as a solution to solve the problem but do not address the hidden consequences. Image classification methods do not match the outstanding performance which these methods have yet to achieve [35].

The approaches discussed in this section begin by determining a point of interest which lets them generate vectors containing essential frame areas. These approaches need to apply motion tracking algorithms for their operation. The research studies achieve satisfactory accuracy but come with a decreased efficiency either through time complexity or other factors. To qualify as a real-time detection method for violence such as real-time video surveillance requires instant crime identification [36].

3. Methodology

This research presents a framework that develops an automatic violence detection system through deep convolutional networks to sort violent and non-violent video frames with five distinct phases shown in Figures 2-4.

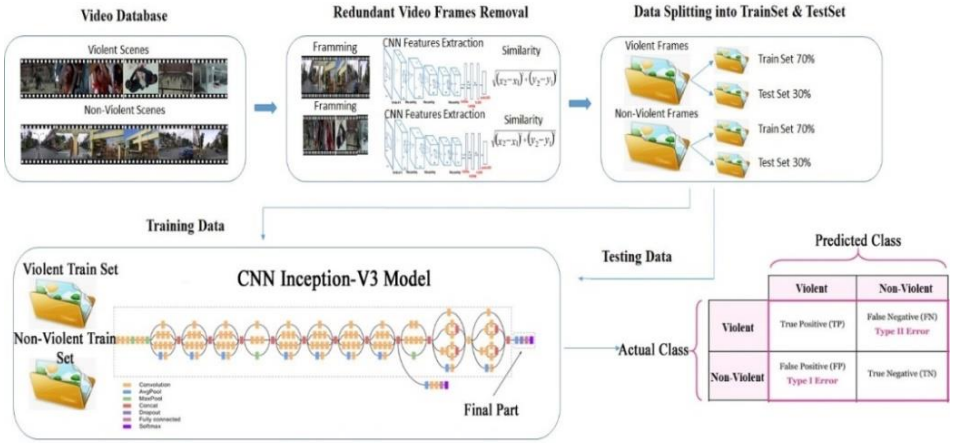

Figure 2. Proposed Framework

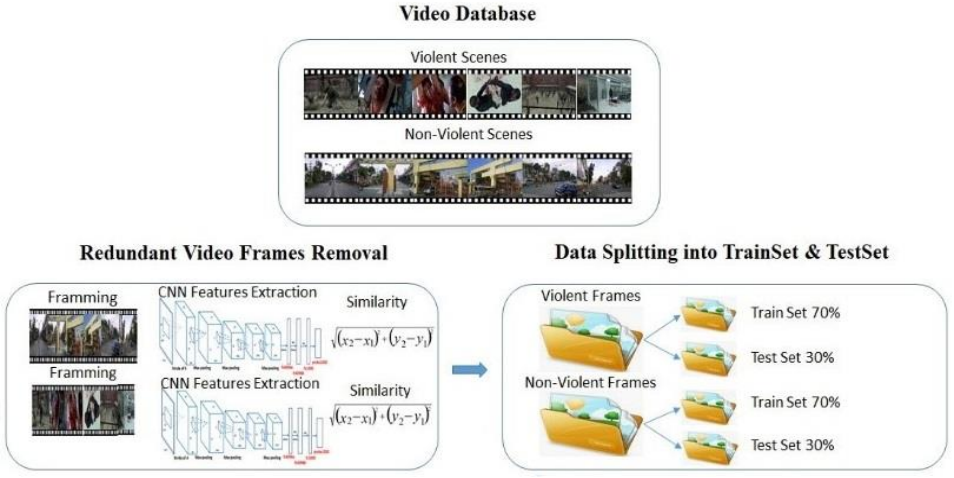

Figure 3. Non-Redundant Video Frames Splitting

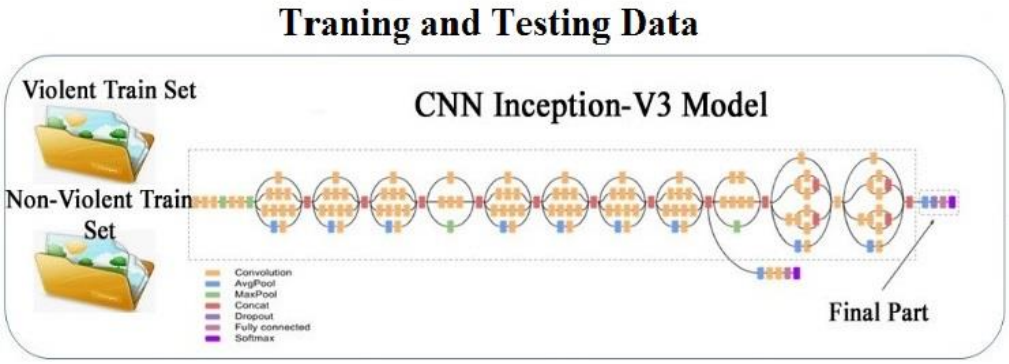

Figure 4 . Inception V3 Model Training and Testing Phase

The system uses a design structure that achieves faster computations alongside maintaining high accuracy levels. A systematic process presented in the diagram operates as an efficient pipeline that enables decision-making with feature learning. The proposed system utilizes the Real-life Violence Situations and Hockey Fight datasets for performing its training and testing operation.

3.1. Acquisition of Datasets

The proposed real-time violence detection method requires evaluation through two selected datasets in this study. These datasets are discussed.

3.1.1. Hockey Fight Dataset

This hockey fight dataset consists of 1000 videos that National Hockey League (NHL) has captured while using one camera which moves while recording. The dataset contains 500 fighting shots and an equal number of non-fighting shots. The clips contain 40 frames which present a resolution of 360 x 288 [18]. The Hockey Fight Dataset contains the clips that are depicted in Figure 5.

Figure 5. Hockey Fight Dataset Clips.

3.1.2. Real-Life Violence Situation Dataset:

The benchmark dataset includes 2000 videos (1000 violent scenes and 1000 normal scenes) that the researchers obtained from street fight videos on online video websites [24]. The videos have a resolution of 1280×720 pixels (24-bit RGB) with a frame rate of 25 FPS and audio frequency of The resolution for all videos amounts to 1280×720 pixels (24-bit RGB) using 25 FPS frame at 44 kHz audio frequency [39]. This dataset serves as a standard reference for violence detection studies particularly when researchers employ oriented violent flows together with 3D Convolutional Neural Networks and spatiotemporal characteristics [37].

The violent situation clips from real life are included in Figure 6.

Figure 6. Real-Life Violence Situation Dataset Clips

3.2. Video Database Preparation

The video database consists of two fundamental sections named Violent Scenes and Non-Violent Scenes to begin the process. Control classes that serve as non-violent scenes exist without physical confrontations or violent interactions or other forms of violence whereas violent scenes display such content. The research uses Real-life Violence Situations and Hockey Fight as its main datasets. The Real-life Violence Situations Dataset provides genuine content depicting real violent situations that include street fights along with security camera and violent confrontation recordings. Hockey Fight Dataset specifically holds filmed hockey matches with their violent events including fights during game play. Different data situations in the datasets allow the system to discover common violent behaviors across multiple environments which enhances its ability to detect violence during practical applications.

3.3. Redundant Video Frames Removal

Video files consist of additional frame images which bring no substantial details while taxing system performance. The video processing starts with dividing it into separate frames as the first step for maximum efficiency. The CNN feature extraction process takes video frames to extract vital visual features from each individual frame. A similarity measurement based on Euclidean distance method determines the redundancy between consecutive video frames. The elimination process of small variation frames results in redundant data reduction which shortens the dataset scale without damaging essential information. By removing repetitive data through this method the training process becomes faster and more efficient and the model becomes better at generalizing information from training datasets.

A computational approach based on learning techniques presents itself as an efficient solution for visual feature extraction within computer vision applications dedicated to video frame reduplication. The identification of Content-Based Feature Retrieval (CBFR) in a video uses various feature detectors and descriptors as part of the examination. SIFT stands as the conventional feature transform method which selects interest points together with extracting features. We recommend using the inception-v3 CNN as the appropriate model to perform feature extraction. The research uses unsupervised and supervised classification algorithms to show variations in their operational results.

The system obtains better precision and efficiency through a similarity measure methodology that determines frame-comparative similarities. A frame remains in the system based on results from the distance calculation. The Content-Based Image Retrieval (CBIR) system detects image repetition by using both features and similarity measures as detection criteria. The video procedure begins with frame conversion before redundancy detection happens based on threshold guidelines. The Inception-v3 feature extraction process demands images to be converted into 299 × 299 RGB format before

continuation. Euclidean and cosine along with Hamming similarity serve as state-of-the-art measure for the system.

3.4. Primary Consideration of similarity

A straight-line measurement of spatial distance known as Euclidean distance establishes the metric through which we define two points within Euclidean space. The transformation of Euclidean space into a metric space became possible through this metric while its induced norm remains the Euclidean norm. The earlier writings occasionally name the measurement as the Pythagorean metric. The abbreviation L2 distance stands for the alternative term of the Euclidean norm also known as the L2 norm. The calculation determines their L2 distance which also corresponds to the L2 norm. The procedure to obtain non-redundant video frames is shown step-by-step in Figure 7. A pair of video frames X, Y meets the similarity requirement when X, Y distance measured by L2 format shows a value of V=15.

$$\text{Feature\_1} = [13, 24, 35, 16, 57, 32, 53, 31, 25, 26, 37] \quad (1)$$

$$\text{Feature\_2} = [26, 12, 17, 5, 25, 16, 11, 22, 36, 21, 17] \quad (2)$$

$$V = \text{Feature\_1} - \text{Feature\_2} \quad (3)$$

$$V\_2 = V * V\ T \quad (4)$$

$$\text{Distance} = (V\_2) \quad (5)$$

Figure 7. Redundant video frames removal using Siamese CNN model.

The second process detects duplicate video frames throughout the database. A large number of video frames duplicate their content in videos playing at high frame-per-second (FPS) rates. The model operates without supervision to extract superfluous video frames from the database access. A learning system based on CNN attributes together with Euclidean distance identifies preferred video frames from the database.

3.5. Data Splitting into Training and Testing Sets

The remaining non-duplicated video frames receive classification as Violent Frames or Non-Violent Frames. The information is split into 70% training material that allows feature acquisition for the model

during training yet the remainder 30% testing content evaluates model performance on unshown data. A hold-out cross-validation technique serves as an appropriate choice for large datasets and complex convolutional models. Among different cross-validation methods the hold-out approach delivers superior generalization capabilities without affecting training efficiency detrimentally.

3.6. Transfer Learning in the Inception-v3 Convolutional Neural Network (CNN)

Transfer learning enables deep learning through model modification of large-dataset trained pre-build networks to suit new related applications. Such an approach reduces training time along with model parameters through optimal enhancements particularly during scenarios with minimized labeled data availability.

3.6.1. Applying Transfer Learning in the Inception-v3 Model

Transfer learning becomes part of the fourth study phase when it is applied to Inception-v3 CNN model. The model performs automatic feature extraction directly from images instead of using human-made features. The pre-trained CNN structure Inception-v3 derives from large datasets including ImageNet for effective feature extraction.

Transfer learning comes into play to classify features that were previously extracted. The process of fine-tuning Inception-v3 model parameters uses a frozen state for all layers preceding the last three layers. General image features arise from early layers of the network while the later layers interpret distinct patterns which appear only in the targeted dataset. Special modifications to the model's last three layers help it acquire new knowledge from distinct classification issues which enhance the system's accuracy while maintaining efficiency.

Model performance evaluation takes place in the concluding stage by measuring accuracy and precision as well as recall and F1-score to determine task efficiency.

3.7. CNN Inception-V3 Model Training

The system uses Deep Convolutional Neural Network (CNN) along with Inception-V3 network as its base structure. The training process consists of the following fundamental elements: 1) Feature Extraction allows the system to draw spatial and temporal elements from video frames to separate violent content from non-violent videos. The Inception-V3 pre-trained model undergoes adaptation through removing obsolete weights as well as tags and prejudices to become focused on identifying violence-specific features independently. The model uses Stochastic Gradient Descent with Momentum (SGDM) optimizer as the optimization method to reach faster convergence rates while preventing local minima trapping.

Small learning rate value of 0.0001 enables stable convergence while preventing sudden weight modifications. Deep learning and transfer learning approaches enable the model to represent features much better and thus enhance its classification abilities.

3.8. Model Testing and Performance Evaluation

Evaluation takes place on the test dataset consisting of 30% of the overall data after the training process ends. The classification results appear in a confusion matrix which displays the precise breakdown of accurate predictions combined with wrong classifications into True Positive (TP) and True Negative (TN) along with False Positive (FP) – Type I Error and False Negative (FN) – Type II Error metrics.

System performance with a lower False Negative Rate allows for proper detection of violent scenes so important incidents can be minimized. The False Positive Rate (FP) system underwent optimization so that it would generate minimum unnecessary alerts.

The work proposed delivers multiple essential contributions: 1) It achieves highest efficient rates for Hockey Fight dataset and Real-life Violence Situations dataset. Data processing operates effectively because the system deletes unnecessary frames which enable better computational efficiency and faster training performance. The system presents strong generalization capabilities due to its independent performance across different datasets despite traditional dataset requirements. The customization of Inception-V3 CNN enables the system to produce better feature patterns which results in better classification outcomes. The security system proves its worth in real-time surveillance applications because it detects violence effectively while excluding non-violent incidents.

The Deep Convolutional Networks-based Automatic Violence Detection System operates as a quick and dependable approach for video violence detection. The model achieves both high accuracy and efficient computation because of its CNN-based features and transfer learning along with similarity measure implementation. This method beats previous state-of-the-art approaches while having potential applications in public security monitoring together with police work and video system analytics.

4. Results And Discussions

4.1. Performance Evaluation of Proposed Model

A description of the utilized performance metrics for comparing the proposed violence detection system appears in this section. Standard classification performance metrics from machine learning practice act as evaluation measures. The detection system uses Accuracy and precision as well as recall and F1-score (F-measure) to evaluate its ability to classify video frames as violent or non-violent.

4.1.1. Performance Metrics

The classification performance is measured using four key metrics, calculated based on the confusion matrix (Figure 8):

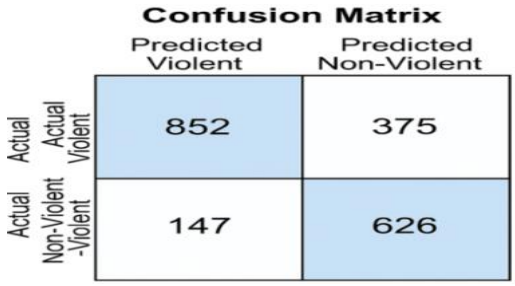

Figure 8. Confusion Matrix Values – Real-Life Violence Dataset

The confusion matrix in Figure 8 consists of values: True Positive (TP): The numbers of correctly (852) identified violent frames as well as False Negative (FN): The number of incorrectly (375) identified violent frames as non-violent but also includes False Positive (FP): The number of incorrectly (147) classified non-violent frames as violent and True Negative (TN): The number of correctly (626) identified non-violent frames as non-violent.

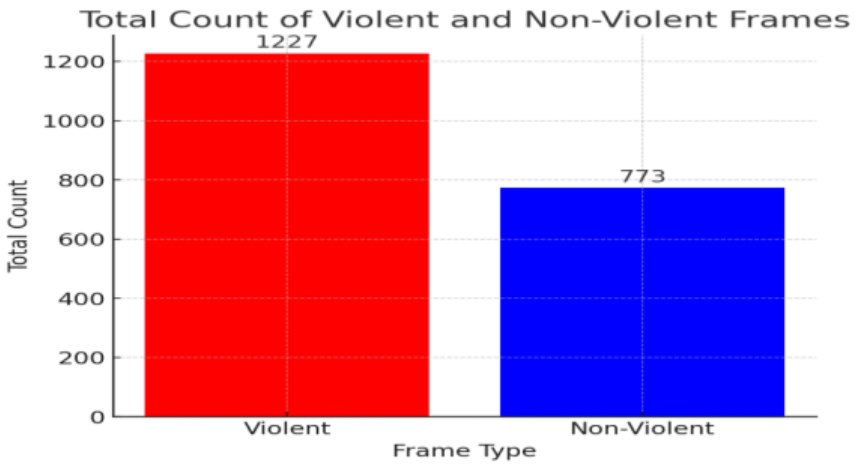

Figure 9. Distribution of Violent and Non-Violent Frames – Real-Life Violence Dataset

The addition of actual cases from confusion matrix produces the total count of violent and non-violent frames in Figure 9. The total number of violent frames in analysis equals the combination of both True Positive (TP = 852) results and False Negative (FN = 375) outcomes resulting in 1227. The data contains a total of 773 frames comprising 626 correct non-violent items and also 147 incorrect non-violent frames. The total number of used frames amounts to 2000 consisting of 1227 violent frames and 773 non-violent frames. According to the displayed bar chart a greater number of violent frame segments exist compared to non-violent ones.

The performance measures become as follows when using these criteria:

Accuracy (AC): The AC measurement evaluates the number of correctly tagged frames from the total frame count.

$$AC = \frac{TP + TN}{TP + FN + TN + FN} \tag{6}$$

Precision (PR): PR measures the true count of violent frames among those labeled as violent.

$$PR = \frac{TP}{TP + FP} \tag{7}$$

Recall (RE) (Sensitivity): The model correctly identifies violent frames through its recall measurement technique (Sensitivity).

$$RE = \frac{TP}{TP + FN} \tag{8}$$

F1-Score (F-Measure): F-Score calculates a compromise between precise assignment and complete recall of violent action frames by using their harmonic mean.

$$F1-Score = 2 \times \frac{PR \times RE}{PR + RE} \tag{9}$$

4.1.2. Confusion Matrix Representation

Figure 10 provides a visual representation of classification results using a confusion matrix. The model's actual classes correspond to the matrix rows at the same time the predicted classes match each column. The proper classifications form a pattern which follows the diagonal yet misclassifications exist in the parts which do not correspond to the diagonal. Dark blues along the confusion matrix represent predictions where the classifier showed no sign of choosing that class whereas yellow areas indicate 50% certainty of certain predictions and dark red shows high confidence rate of 100% for identifying that class.

The best possible classification appears as dark red cells along the confusion matrix main diagonal while no misclassifications should be blue colored at off-diagonal positions.

4.1.3. Accuracy and Error Rate Analysis

The evaluation includes a report of the top-1 error rate together with total accuracy assessment. The measurement indicates the weakest class performance through which researchers detect areas of model inadequacy. Networks can reveal unseen patterns by using graphs and confusion matrices while they would also compare these results to distinguish between various trends.

4.1.4. Training Progress and Convergence

A graphical representation (Figure 10) shows accuracy and loss levels during every 40 training iteration cycle for examining model learning patterns. Visual analysis becomes feasible through convergence patterns which exhibit learning efficiency and speed as well as pre-trained model effects on learning capabilities and accuracy outcomes.

Through this testing protocol the proposed system obtains full analysis and comparison testing to deliver optimal parameters and improved classification performance.

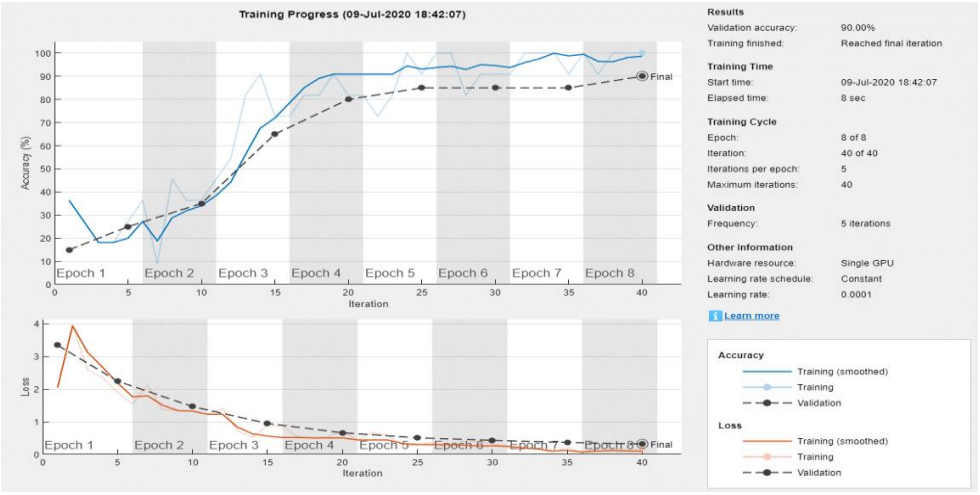

Figure 10. Training Progress accuracy and loss with 40 Iterations.

4.2. Analysis and Discussion of Validation

4.2.1. Importance of Validation in Deep Learning

The model needs validation to achieve generalization along with appropriate performance when processing new data. During the third phase of this research the dataset is divided into training and testing sections through hold-out cross-validation because this proves effective for big image databases and costly convolutional neural network (CNN) models.

4.2.2. Choice of Hold-Out Cross-Validation

The researchers choose the hold-out method because it demonstrates high efficiency when working with extensive and complex image datasets. Among other validation methods the hold-out method stands out because it delivers fast assessments with decreased training periods.

The 70%-30% training-testing split offers an adequate training model capacity combined with sufficient assessment material to guarantee accurate results. In deep learning research this particular split arrangement has gained universal acceptance since it delivers proper training effectiveness and accurate testing results.

4.2.3. Model Training Parameters

Table 1 shows critical training performance aspects for which the model receives training for 50 epochs followed by 100 epochs to track various training duration effects. More training cycles help extract advanced characteristics yet they increase the possibility of model misfitting. Distribution of 64 batch items proves best for achieving both computational performance and steady training convergence. The model relies on a small learning rate (0.0001) that produces stable convergence because it stops disruptive updates from occurring. The optimized SGDM algorithm serves to speed up the convergence rate while overcoming local minima problems especially relevant to deep learning algorithms. Several toolboxes named DIP, CV, and DL within MATLAB indicate an advanced computational system that supports different methods of image processing alongside learning techniques.

Table 1. Key Training Parameters and Methods

Parameters	Values/Methods
Epochs	50 and 100
Batch Size	64
Learning Rate	0.0001
Optimization	SGDM
Toolboxes	DIP, CV, and DL
Datasets	Real-Life Violence Situations, Hockey Fight
Similarity Measure	Euclidean Distance
Cross Validation	70/30 Percent

4.2.4. Datasets and Application

The analysis relies upon the Real-life Violence Situations and Hockey Fight datasets. The available datasets feature labeled violent and non-violent events which make these collections appropriate for building video-based violence detection systems. Such wide and realistic datasets during training improve system performance across multiple operational conditions.

4.2.5. Similarity Measure and Evaluation

The Euclidean distance functions as the selected similarity assessment method. This measure allows researchers to determine vector closeness thus identifying superfluous frames and validating prediction outcomes. The use of Euclidean distance presents limitations because it reacts to change in scale so researchers should evaluate alternatives such as cosine similarity.

4.2.6. Performance Considerations

Computational Efficiency increases because the large dataset and CNN model benefit from the hold-out method which reduces computational expenses. The use of SGDM optimizer along with a moderate batch size and learning rate contributes to maintaining a proper control over overfitting. The Feature Extraction & Similarity Matching process benefits from the Euclidean distance measure although implementing extra techniques could enhance its accuracy in frame differentiation.

The validation strategy provides a structured methodology to manage big data processing with efficient overfitting reduction capabilities. The robustness potential of the method could be optimized through using stratified k-fold cross-validation with additional similarity measures when computational limits permit. The performance can be optimally refined through additional parameters adjustments which include epochs, learning rate and optimizer selection.

4.3. Real-Life Violence Dataset Analysis

Two training epochs were applied to evaluate the Real-Life Violence dataset with 50 and 100 as options. An increase in epochs resulted in a moderate improvement of performance metrics contained in Tables 2 and 3. The Training Parameters include Epochs set to 50 and 100 combined with Batch Size at 64 and the use of SGDM as the optimizer.

Table 2. Confusion Matrix Values – Real-Life Violence Dataset

Epoch	Class	Violent (%)	Non-Violent (%)
50	**Violent**	85.27	14.73
	Non-Violent	37.52	62.48
100	**Violent**	80.31	19.69
	Non-Violent	23.09	76.91

Table 3. Performance Metrics – Real-Life Violence Dataset

Metric	50 Epochs	100 Epochs
Accuracy	73.87%	78.61%
Precision	85.27%	80.31%
Recall	69.44%	77.66%
F-Measure	76.54%	79.21%

43.1. Confusion Matrix Analysis

During 50 epochs the model reached 73.87% accuracy while correctly identifying 85.27% violent videos while falsely identifying 37.52% non-violent videos as violent. A total of 100 epochs increased accuracy to 78.61% while recall became 77.66% and the classification rate of non-violent videos reached 76.91%. The loss of precision went from 85.27% to 80.31%, showing a minor disadvantage to both true and false detections.

4.3.2. Performance Metrics Comparison

The classification performance improved as accuracy values rose from 73.87% to 78.61%. The detection of violent instances improved through a better recall rate. The precision rate decreased which demonstrated that false positive cases increased slightly. The performance evaluation through F-measure showed increased results from 76.54 to 79.21 as an indicator of balanced system improvement.

4.4. Hockey Fight Dataset Analysis

The training method led to performance assessments that appeared in Table 4 (confusion matrix) and Table 5 (performance metrics) for Hockey Fight Dataset. The SGDM optimizer operated with Batch Size = 64 for 50 to 100 epochs.

Table 4. Confusion Matrix Values – Hockey Fight Dataset

Epoch	Class	Violent (%)	Non-Violent (%)
50	**Violent**	73.48	26.52
	Non-Violent	38.75	61.25
100	**Violent**	85.27	14.73
	Non-Violent	37.55	62.45

Table 5. Performance Metrics – Hockey Fight Dataset

Metric	Epoch 50	Epoch 100
Accuracy	67.36	73.86
Precision	73.48	85.27
Recall	65.47	69.42
F-Measure	69.24	76.53

4.4.1. Confusion Matrix Analysis

Throughout 50 epochs the model demonstrated 67.36% accuracy which correctly identified 73.48% violent videos yet it wrongly labeled 38.75% non-violent videos as violent. The model reached 73.86% accuracy after 100 epochs while violent video classification increased to 85.27% but the rate of misclassifying non-violent videos stayed comparable..

4.4.2. Performance Metrics Comparison

The Hockey Fight dataset exhibited performance patterns similar to Real-Life Violence but produced less significant achievement results. Precision decreased mildly when compared to the steady improvement of accuracy, recall and F-measure.

4.5.performance Comparison of both Datasets

The performance metrics for both datasets are depicted in Figure 11. Key observations include:

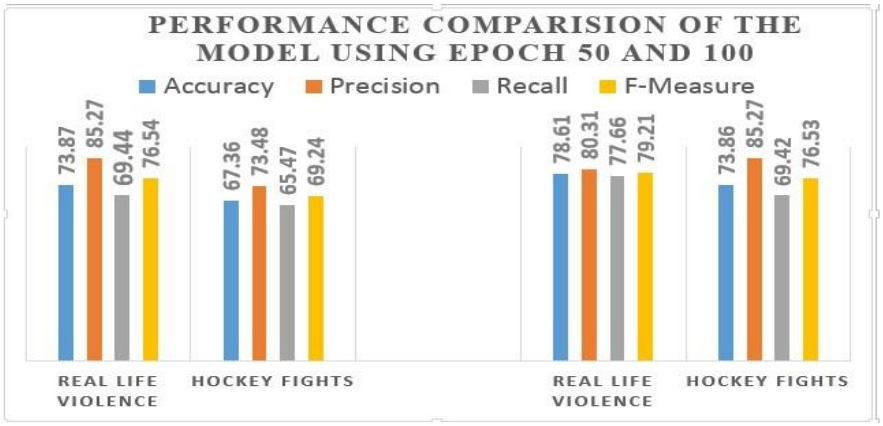

Figure 11. Performance evaluation of both datasets in Bar-chart

The performance metrics of accuracy and recall increase when the epoch parameter rises from 50 to 100 across both datasets. The Real-Life Violence dataset gained superior results from additional epochs than the Hockey Fight dataset. The varying performance enhancements stem from differences between video quality and illumination levels and camera panning in the datasets. The Hockey Fight dataset includes moving camera videos that cause both motion blur effects and changes in camera angle throughout the video sequence. The Real-Life Violence dataset includes static videos recorded inside and outside so the model can easily recognize recurring patterns.

The study reached better classification results by extending epoch count but this strategy might not universally apply to various datasets. Applications of this method work best for stationary videos yet require additional enhancement to sort multiple video movements. Future developers should analyze alternative models and extra classification groups as well as feature extraction strategies to optimize prediction accuracy.

4.6. Comparative Analysis of the Proposed Model with Existing Research

Model effectiveness in violence detection relies upon the combination of data sets and extraction methods together with identified classification techniques. This section performs an extensive evaluation between our proposed methodology and existing leading models documented in research literature. Table 6 shows precision statistics from various techniques while using different datasets.

Table 6. Comparison of Proposed Model with Existing Research work

Author	Dataset	Techniques	Accuracy
Bermejo Nievas, et al. [18]	Hockey Fight, Movie violence	Bag-of-Words + STIP, and Mo SIFT.	89.5%, 70.00%
Zhang, Tao, et al. [38]	Hockey Fight, Real Life violence	SVM + Adaboost	67.50%, 68.00%
Gao, Yuan, et al. [1]	Hockey Fight, Behave	2D CNN, Hough Forest	64.6%, 71.4%
Serrano, Ismael et al. [39]	Hockey Fight, Movie Violence	SVM	70.0%, 89.5%
Serrano Gracia, et al. [40]	Hockey Fight, Movie Violence	Random Forest	70%, 90%
Soliman, Mohamed Mostafa, et al. [41]	Hockey Fight, Real Life Violence Situation.	CNN (Transfer Learning)	52.2%, 68.2%
Hsairi, Lobna, et al. [42]	8-class annotated dataset (violent/non-violent)	CNN architectures (InceptionV3, MobileNetV2,	71%

		ResNet-152V2, VGG-16)	
de Andrade, et al. [43]	AIRTLab and PA-100K. The AIRTLab dataset	CNNs	73%,
Proposed Model	Hockey Fight Real Life Violence Situation	Siamese + CNN	77.01%, 78.16%

The proposed method uses CNN and Siamese architectures to perform robust extraction and classification of features. The deep learning method employs CNN with Siamese structures to extract features because it outperforms traditional handcrafted components consisting of BoW and SIFT as well as machine learning approaches using SVM and AdaBoost. Multiple datasets present an obstacle to generalization for the feature representation achieved by 2D CNN and InceptionV3 models which operate from a CNN framework.

4.6.1. Dataset and Generalization Issues in Existing Methods

The current techniques for feature extraction depend on manually designed approaches including Bag-of-Words model and Space-Time Interest Points (STIP) and Motion SIFT (Mo SIFT). The research by Bermejo Nievas et al. [18] obtained 89.5% accuracy for Hockey Fight dataset yet their method showed 70% performance in the Movie Violence dataset. The research by Zhang et al. [38] presented two classifier techniques using Support Vector Machines (SVM) and AdaBoost while their performance showed weak generalization capability with accuracy levels at 67.5% and 68%.

Gao et al. [1] integrated 2D CNNs with Hough Forest in their method which resulted in limited accuracy at 64.6% and 71.4% because previous CNN architectures had poor feature extraction capabilities. Serrano et al. [39] teamed up with Serrano Gracia et al. [40] to apply SVM with Random Forest classifiers for processing the Movie Violence dataset which delivered 89.5% and 90% performance marks yet both performed poorly with 70% accuracy on the Hockey Fight dataset.

Soliman et al. [41] applied transfer learning systems with CNNs which resulted in sub-optimal results of 52.2% and 68.2% accuracy because the direct transfer of learning from unrelated domains does not always produce effective outcomes. Hsairi et al. [42] examined InceptionV3, MobileNetV2, ResNet-152V2, and VGG-16 CNN structures on an 8-class dataset yielding 71% accuracy yet the study had no explicit method for eliminating superfluous video frames which could possibly reduce operational efficiency. Research from de Andrade et al. [43] used CNNs to analyze AIRTLab along with PA-100K datasets giving 73% accuracy results in their evaluation however their work failed to focus on authentic violence detection applications..

4.6.2. Strengths and Improvements of the Proposed Model

Our proposed approach merges Siamese networks and CNN to greatly improve feature extraction and generalization. The detection of similarity is enhanced through the architecture of the Siamese network, allowing better distinction between violent and nonviolent frames. InceptionV3 is fine-tuned with the model by removing its last three layers and using a customized feature extraction approach. A hold-out cross-validation method (70% for training, 30% for testing) is employed to ensure strict testing.

Compared to the state of the art, our model achieves 77.01% and 78.16% accuracy on the Hockey Fight and Real-Life Violence Situation datasets, respectively. These accuracies represent an improved balance between dataset generalization and feature learning than traditional methods such as SVM, AdaBoost, and baseline CNN classifiers. In addition, removal of redundant frames improves training efficiency by reducing computational overhead while maintaining classification accuracy.

4.6.3. Key Findings of the Proposed Model

The key results of this comparative analysis are that the proposed Siamese + CNN model outperforms traditional and state-of-the-art methods with accuracy rates of 77.01% and 78.16% on the Hockey Fight and Real-Life Violence Situation datasets. By refreshing the inception-v3 CNN model by removing old weights, labels, features, and biases, the system extracts more significant features, enhancing classification performance. Unlike other models that have performance variation across datasets, the proposed model is stable and has high generalization capability. Removing redundant frames makes the learning process easy, achieving maximum speed and efficiency in model learning. The use of CNN and Siamese architecture enhances contrastive learning by increasing the distinction between violent and non-violent scenes. Usage of Content-Based Image Retrieval (CBIR) techniques ensures redundant frames are eliminated; subsequently ensuring training and validation are optimal. These findings confirm the improved accuracy, efficiency, and robustness of the proposed model in violence detection.

4.6.4. Research Contributions of the Proposed Model

The contributions of research are expressed in a number of ways. First, the method proposed improves frame comparison, and hence classification accuracy is improved. Second, the training is rendered efficient by removing redundant frames, thus minimizing computational overhead. Thirdly, with the adaptation of the inception-v3 model, improved feature representations are learned and classification performance is achieved.

The model is also evaluated on benchmark violence detection datasets to ensure efficiency and reliability. Whereas other models are specific to a dataset, the proposed model shows uniform performance for any dataset and is therefore more versatile.

4.6.5. Model Evaluation and Performance Assessment

The evaluation process is performed in a controlled way, and the results are reproducible and consistent. The model is trained and tested with widely used violence detection datasets to provide an unbiased comparison with existing models. The inception-v3 model is preprocessed by removing previous weights and biases, increasing its feature extraction ability. A 70/30 hold-out cross-validation procedure yields an objective assessment of model performance. The model is compared with the state of the art methods and demonstrates to perform better than them in classifying violence. Simulation of the redundant frames' removal greatly accelerates the training, decreasing the computational load at no loss in accuracy.

The suggested Siamese + CNN-based framework offers a new and effective way of detecting violence from videos. Equipped with the ability to eliminate the disadvantage in current algorithms, including the lack of stability in accuracy and ineffective processing of frames, the suggested system proves to be a highly effective solution for real applications in security and surveillance.

5. Conclusion

Real-time violence detection problem in surveillance videos with urgency was targeted in this study through the use of a deep learning technique that used the Inception-v3 model and transfer learning. The overarching theme of the study was overcoming shortcomings in the form of redundant frames, set dependency, and insufficiency of generalization in conventional methods. Accurate contributions reported that the proposed system achieved 78.61% and 73.86% accuracies on Real-Life Violence and Hockey Fight databases, respectively, and outperformed state-of-the-art. Euclidean distance greatly enhanced computational efficiency by eliminating redundant frames. Transfer learning enhanced feature extraction and classification efficiency. The potential of such findings is wide for practical applications in real-world scenarios such as public monitoring, law enforcement, and web filtering, where accurate

and timely detection of violence is essential. The research adds by presenting a resilient model that combines Siamese networks and CNNs with better generalization and efficiency, validated on heterogeneous sets of data. Drawbacks are vulnerability to motion blurring in motion sequences and use of pre-specified threshold values for redundancy elimination. Future work can include using adaptive thresholding methods, integration of temporal models such as LSTMs to enhance motion analysis, and multi-class violence detection generalization. Otherwise, the experiment on larger and more diverse datasets would also prove the scalability of the model.

This work promotes automatic detection of violence using a consistent, efficient, and viable solution. In its encounter with present-day problems and positing where to go next, it presents a window of opportunity for more evolved and dynamic systems in security and video monitoring.

References

1. Gao, Yuan, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. "Violence detection using oriented violent flows." Image and vision computing 48 (2016): 37-41.
2. Eneim, Maryam. "An Intelligent Method for Violence Detection in Live Video Feeds." PhD diss., Florida Atlantic University, 2016.
3. Febin, I. P., K. Jayasree, and Preetha Theresa Joy. "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm." Pattern Analysis and Applications 23, no. 2 (2020): 611-623.
4. Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., vol. 3, pp. 32-36. IEEE, 2004.
5. Gong, Faming, Chuantao Li, Wenjuan Gong, Xin Li, Xiangbing Yuan, Yuhui Ma, and Tao Song. "A Real-Time Fire Detection Method from Video with Multifeature Fusion." Computational intelligence and neuroscience 2019, no. 1 (2019): 1939171.
6. Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." The American Statistician 46, no. 3 (1992): 175-185.
7. Yu, Jing, Wei Song, Guozhu Zhou, and Jian-jun Hou. "Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation." Multimedia Tools and Applications 78 (2019): 8497-8512.
8. Zeb, Palwasha, Qasim Arbab, Muhammad Qasim Khan, and Haider Ali. "Classification of Acute Myeloid Leukaemia using deep learning features." The Sciencetech 4, no. 1 (2023).
9. Arbab, Qasim, Muhammad Qasim Khan, and Haider Ali. "Automatic Detection and Classification of Acute Lymphoblastic Leukemia Using Convolution Neural Network." The Sciencetech 3, no. 4 (2022).
10. Khan, Muhammad Qasim, Steinar Hidle Andresen, and Muhammad Inamul Inam Ul Haq. "Handover architectures for heterogeneous networks using the media independent information handover (mih)." Computing and Informatics 35, no. 1 (2016): 177-202.
11. Hussain, Shah, and Muhammad Qasim Khan. "Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning." Annals of data science 10, no. 3 (2023): 637-655.
12. Ullah, Kifayat, and Muhammad Qasim. "Google stock prices prediction using deep learning." In 2020 IEEE 10th international conference on system engineering and technology (ICSET), pp. 108-113. IEEE, 2020.
13. Khan, Muhammad Qasim. "Signaling storm problems in 3gpp mobile broadband networks, causes and possible solutions: A review." In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 183-188. IEEE, 2018.
14. Wafa, Rubab, Muhammad Qasim Khan, Fazal Malik, Akmalbek Bobomirzaevich Abdusalomov, Young Im Cho, and Roman Odarchenko. "The impact of agile methodology on project success, with a moderating role of Person's job fit in the IT industry of Pakistan." Applied Sciences 12, no. 21 (2022): 10698.
15. Breiman, Leo. "Bagging predictors." Machine learning 24 (1996): 123-140.
16. Khan, SanaUllah, Naveed Islam, Zahoor Jan, Ikram Ud Din, and Joel JP C. Rodrigues. "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning." Pattern Recognition Letters 125 (2019): 1-6.
17. Saber, Abeer, Samar Elbedwehy, Wael A. Awad, and Esraa Hassan. "An optimized ensemble model based on meta-heuristic algorithms for effective detection and classification of breast tumors." Neural Computing and Applications 37, no. 6 (2025): 4881-4894.

18. Bermejo Nievas, Enrique, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. "Violence detection in video using computer vision techniques." In Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14, pp. 332-339. Springer Berlin Heidelberg, 2011.
19. Itcher, Yossi. Real-Time Detection of Violent Crowd Behavior. Open University of Israel, 2013.
20. Chen, Ming-yu, and Alex Hauptmann. "Mosift: Recognizing human actions in surveillance videos." Computer Science Department 929 (2009).
21. Cheng, Guangchun, Yiwen Wan, Abdullah N. Saudagar, Kamesh Namuduri, and Bill P. Buckles. "Advances in human action recognition: A survey." arXiv preprint arXiv:1501.05964 (2015).
22. Wang, Heng, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. "Dense trajectories and motion boundary descriptors for action recognition." International journal of computer vision 103 (2013): 60-79.
23. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20 (1995): 273-297.
24. Hassner, Tal, Yossi Itcher, and Orit Kliper-Gross. "Violent flows: Real-time detection of violent crowd behavior." In 2012 IEEE computer society conference on computer vision and pattern recognition workshops, pp. 1-6. IEEE, 2012.
25. Freund, Yoav, Robert Schapire, and Naoki Abe. "A short introduction to boosting." Journal-Japanese Society For Artificial Intelligence 14, no. 771-780 (1999): 1612.
26. Malik, Fazal, Muhammad Suliman, Shehla Shaha, Muhammad Qasim Khan, and Abd Ur Rub. "Optimizing Pneumonia Diagnosis during COVID-19: Utilizing Random Forest for Accurate Classification and Effective Public Health Interventions." Journal of Computing & Biomedical Informatics 7, no. 01 (2024): 297-312.
27. Malik, Fazal, Muhammad Suliman, Muhammad Qasim Khan, Noor Rahman, and Mohammad Khan. "Optimized XGBoost-based model for accurate detection and classification of COVID-19 pneumonia." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).
28. Suliman, Muhammad, Fazal Malik, Muhammad Qasim Khan, Ashraf Ullah, Noor Rahman, and Said Khalid Shah. "A Convolutional Neural Network (CNN) Based Framework for Enhanced Diagnosis and Classification of COVID-19 Pneumonia." VAWKUM Transactions on Computer Sciences 12, no. 2 (2024): 220-240.
29. Suliman, Muhammad, Fazal Malik, Muhammad Qasim Khan, Irfan Ullah, and Abd Ur Rub. "Integrating data augmentation with AdaBoost for effective COVID-19 pneumonia classification." Journal of Computing & Biomedical Informatics 7, no. 01 (2024): 590-605.
30. Khan, Muhammad Qasim, Fazal Malik, and Noor Rahman. "Optimized Sentiment Classification of Google Play Store App Ratings Using Advanced Machine Learning Models." VFAST Transactions on Software Engineering 12, no. 4 (2024): 252-266.
31. Shah, Masroor, Fazal Malik, Muhammad Suliman, Noor Rahman, Irfan Ullah, Sana Ullah, Romaan Khan, and Salman Alam. "Dark Data in Accident Prediction: Using AdaBoost and Random Forest for Improved Accuracy." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).
32. Malik, Fazal, Muhammad Suliman, Muhammad Qasim Khan, Noor Rahman, Khairullah Khan, and Muhammad Khan. "Optimizing malicious website detection with the XGBoost machine learning approach." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).
33. Malik, Fazal Malik Fazal, Muhammad Suliman Suliman, Irfan ullah Irfan, Shehla Shah Shehla, and Asiya Bibi Asiya. "Enhancing Cyber Security: A Holistic Strategy for Advanced Malicious Website Prediction Using AdaBoost Algorithm." Lahore Garrison University Research Journal of Computer Science and Information Technology 8, no. 3 (2024).
34. Malik, F., A. U. Rahman, A. Ullah, R. Hussain, M. Javed, & S. Ullah. (2024). Optimizing Malicious Website Detection Through Comparative Analysis of Machine Learning Techniques. Pakistan

Journal of Scientific Research, 4(1(Suppl.), 147–161. https://doi.org/10.57041/vol4iss1(Suppl.)pp147-16.
35. Vrigkas, Michalis, Christophoros Nikou, and Ioannis A. Kakadiaris. "A review of human activity recognition methods." Frontiers in Robotics and AI 2 (2015): 28.
36. Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." In Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9, pp. 404-417. Springer Berlin Heidelberg, 2006.
37. Ullah, Fath U. Min, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. "Violence detection using spatiotemporal features with 3D convolutional neural network." Sensors 19, no. 11 (2019): 2472.
38. Zhang, Tao, Zhijie Yang, Wenjing Jia, Baoqing Yang, Jie Yang, and Xiangjian He. "A new method for violence detection in surveillance scenes." Multimedia Tools and Applications 75 (2016): 7327-7349.
39. Serrano, Ismael, Oscar Deniz, Jose Luis Espinosa-Aranda, and Gloria Bueno. "Fight recognition in video using hough forests and 2D convolutional neural network." IEEE Transactions on Image Processing 27, no. 10 (2018): 4787-4797.
40. Serrano Gracia, Ismael, Oscar Deniz Suarez, Gloria Bueno Garcia, and Tae-Kyun Kim. "Fast fight detection." PloS one 10, no. 4 (2015): e0120448.
41. Soliman, Mohamed Mostafa, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. "Violence recognition from videos using deep learning techniques." In 2019 ninth international conference on intelligent computing and information systems (ICICIS), pp. 80-85. IEEE, 2019.
42. Hsairi, Lobna, Sara Matar Alosaimi, and Ghada Abdulkareem Alharaz. "Violence Detection Using Deep Learning." Arabian Journal for Science and Engineering (2024): 1-11.
43. de Andrade, João Pedro Freire and Si, Tapas and Nascimento, André C. A. and Cavalcanti, Ana Paula and Miranda, Pericles B. C., Susan: A Deep Learning-Based Architecture for Violence Detection Against Women in Surveillance Videos, 2024. Available at SSRN: https://ssrn.com/abstract=4916438.