IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

NIRBHAY: An AI-Powered Multimodal Threat Detection System for Real-Time Surveillance

Sanjyot Pawar, Talat Kudsi, Yami Bhagat, Rushikesh Jagtap, Rasika Kulkarni
[1]Department of Computer Science, Fergusson College, Pune, Maharashtra, India.
[2]Department of Data Science, Fergusson College, Pune, Maharashtra, India.
[1]rasika.kulkarni@fergusson.edu

Abstract: *Traditional CCTV-based surveillance systems suffer from significant limitations, including dependence on continuous human oversight, operator fatigue, delayed incident response, and frequent missed events. With the rise in violent crimes, weapon-related incidents, and unintentional fire outbreaks, there is an urgent need for automated, real-time threat detection systems. This paper introduces NIRBHAY, an intelligent multimodal surveillance framework that integrates three specialized deep learning modules for comprehensive threat detection.*

The system combines violence detection, weapon recognition, and fire-smoke identification into a unified surveillance pipeline.

For violence detection, we employ a hybrid MobileNetV2–BiLSTM architecture, achieving an F1-score of 0.892, accuracy of 0.893, precision of 0.881, and recall of 0.905. Weapon detection utilizes YOLOv8, attaining a mean Average Precision (mAP50) of 0.927, with class-wise precision of 0.904 for handguns and 0.938 for knives. Fire and smoke detection is implemented using an enhanced YOLOv11 model, achieving mAP50 of 0.941, with fire detection accuracy of 0.958 and smoke detection accuracy of 0.924. The integrated system operates at 25–30 frames per second, enabling real-time performance suitable for live CCTV deployment.

NIRBHAY demonstrates robust and reliable threat detection across multiple threat categories, making it highly suitable for deployment in safety-critical environments such as public buildings, transportation hubs, and commercial infrastructures. The system's ability to process multiple threat types simultaneously while maintaining real-time performance represents a significant advancement in autonomous surveillance technology.

Keywords: threat detection, YOLOv8, YOLOv11, violence detection, weapon detection, fire and smoke detection, MobileNetV2, BiLSTM, deep learning, autonomous monitoring, multimodal surveillance.

I. INTRODUCTION

The rising incidence of violent occurrences, crimes involving weapons, and unintentional fire outbreaks in both public and private settings has made safety and security a critical concern worldwide [8], [10]. Today, surveillance cameras are widely deployed across residential neighbourhoods, commercial complexes, educational institutions, and transportation hubs to monitor such environments [9]. However, continuous human monitoring of multiple video streams remains highly challenging due to human error, fatigue, and delayed decision-making [5]. Any delay in identifying a potential threat can result in severe consequences, including loss of life, physical injury, and property damage [1].

To overcome these limitations, advanced intelligent surveillance systems are required to automate the detection of potentially hazardous events [8]. Recent research has demonstrated the effectiveness of deep learning–based approaches for violence detection using temporal and spatial features extracted from video streams [4], [5], as well as the use of comprehensive surveillance datasets to improve model generalization [2]. In parallel, object detection models such as YOLOv8 have shown strong performance in real-time weapon detection scenarios [1], [6], [9], while lightweight and edge□optimized models enable deployment in resource□constrained environments [3].

NIRBHAY is a real-time AI□driven safety monitoring system designed to address these challenges by enabling automatic detection of violence, weapon presence, and fire or smoke hazards. The system leverages deep learning architectures trained on publicly available datasets and real-world CCTV surveillance footage [2], allowing accurate threat recognition in dynamic environments. Fire and smoke detection capabilities are inspired by recent advances in deep learning–based environmental hazard detection [6], [11], while small□object detection techniques further enhance the system's ability to detect weapons in low□resolution surveillance footage [12].

NIRBHAY continuously monitors live video streams and promptly identifies unusual or dangerous activities. Upon detecting a threat, the system generates alerts to facilitate rapid response and intervention by security personnel. By integrating Convolutional Neural Networks (CNNs), Bidirectional Long Short□Term Memory (Bi□LSTM) networks, and YOLO□based object detection frameworks, the system effectively captures both spatial and temporal information while maintaining real□time performance [1], [4], [9]. The combination of these advanced techniques enhances situational awareness, reduces dependence on constant human supervision, and contributes to safer public spaces through early threat prevention and efficient incident management [3], [8].

II. LITERATURE REVIEW

A substantial body of research has focused on the development of intelligent surveillance systems using computer vision and deep learning techniques. Most existing studies concentrate on single□modality threat detection, such as weapon identification, violence recognition, or fire–smoke detection, rather than providing a unified safety monitoring solution. This section reviews key contributions from recent literature and highlights their strengths and limitations.

Bhatti M. T. et al. [1] proposed a real□time weapon detection system based on YOLOv4, in which a custom dataset comprising multiple weapon categories was created due to the lack of standardized CCTV□based datasets. The model achieved a high F1□score and demonstrated strong precision for pistol detection. However, the system struggled under real□world conditions involving occlusions, poor lighting, and varying camera angles, limiting its robustness. Addressing generalization challenges, Kaya V. et al. [2] introduced a multi□weapon classification framework using deep CNN architectures such as VGG19. Although the approach achieved higher accuracy across seven weapon categories, it required further validation in complex CCTV environments with dynamic backgrounds and real□time constraints.

To improve real□time performance, Ahmed S. et al. [3] developed an optimized Scaled□YOLOv4 model suitable for edge deployment on devices such as Jetson Nano. While the system achieved high accuracy and frame rates, its scalability for large□scale surveillance systems and robustness under diverse illumination conditions were not thoroughly evaluated. Bhatt A. et al. [4] proposed a DenseNet□121□based pipeline for detecting dangerous and explosive weapons, offering a more comprehensive weapon recognition approach. Despite achieving strong classification results, the absence of testing on real□world CCTV footage limited its practical applicability.

Golande R. et al. [5] explored a hybrid threat detection system combining CNN□based classifiers with YOLO□based object detection to reduce false positives in crowded scenes. Although promising, the study lacked clarity regarding dataset specifications and did not fully address partial occlusion issues. For violence detection, Inayathulla M. et al. [6] proposed a hybrid architecture integrating ResNet50V2 with Bi□GRU and Bi□LSTM networks, demonstrating effective recognition of violent activities. However, the work did not sufficiently explore scalability and feasibility for deployment in large□scale or edge□based surveillance infrastructures.

Leng J. et al. [7] introduced a dual□space representation learning approach for violence detection by combining Euclidean and hyperbolic embeddings. While this method improved recognition in ambiguous scenarios, the increased model complexity and computational overhead posed challenges for real□time inference. Similarly, Khan M. Q. et al. [8] proposed a violence detection system based on Inception□V3 with transfer learning, achieving solid performance

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

on benchmark datasets. Nonetheless, the model lagged more recent approaches and lacked extensive temporal modeling capabilities.

Shirole B. S. et al. [9] presented a deep-feature-based violence recognition system that achieved high accuracy on controlled datasets. However, concerns regarding generalization arose due to limited dataset diversity and the absence of real surveillance footage. Mukto M. M. et al. [10] proposed an integrated multi-task surveillance framework combining YOLOv5, MobileNetV2, and face recognition for criminal monitoring. Although the system successfully integrated multiple modules, scalability challenges persisted, and weapon detection accuracy remained comparatively low.

Advancements in small-object detection were explored by Quyyum M. E. et al. [11], who enhanced YOLOv3 with an additional prediction layer to better detect small weapons such as knives and handguns. Despite improved detection, the approach faced trade-offs between speed and accuracy, particularly with low-resolution surveillance data. Vijayakumar K. P. et al. [12] proposed a hybrid YOLOv4 and Faster R-CNN model for multi-weapon detection, achieving higher precision but suffering from computational inefficiency that limited real-time applicability.

Lightweight architectures were investigated by Vijeikis R. et al. [13], who employed a CNN–LSTM framework for violence detection in resource-constrained environments. While suitable for edge deployment, the model's accuracy remained below state-of-the-art solutions. Choqueluque-Roman D. et al. [14] addressed annotation cost challenges using weakly supervised learning techniques, though accurate localization of violent events remained difficult.

In the domain of fire and smoke detection, Alrubaian A. et al. [15] utilized YOLOv5 with enhanced feature extraction for early fire detection. Although the model achieved promising accuracy and speed, smoke detection under low-light conditions remained challenging. Sharma A. et al. [16] introduced FireNet++, a CNN-based model employing multi-scale features to detect fire regions; however, flame-like reflections frequently resulted in false positives.

Rahman A. et al. [17] proposed a cross-domain firearm detection approach by fine-tuning YOLOv7 on diverse datasets, improving robustness across environments. Nevertheless, increased input resolution significantly reduced inference speed. Chen Y. et al. [18] leveraged Graph Neural Networks (GNNs) to model human joint relationships for aggression detection, achieving improved performance in crowded scenes but requiring computationally intensive skeleton extraction preprocessing.

Finally, Patil S. et al. [19] introduced a dual-task YOLO-based framework capable of simultaneously detecting fire and weapons. While the approach improved efficiency by sharing computations, the multi-output architecture increased training complexity.

Overall, existing literature demonstrates substantial progress in individual threat detection modalities—violence, weapons, and fire-smoke detection. However, very few studies integrate all three capabilities into a unified, real-time surveillance system, highlighting a clear research gap that motivates the proposed work.

III. METHODOLOGY

The proposed NIRBHAY system follows a unified real-time surveillance pipeline designed to simultaneously detect violent activities, weapon presence, and fire–smoke hazards from continuous video streams. The architecture consists of three parallel detection modules—Violence Detection, Weapon Detection, and Fire-Smoke Detection—which operate concurrently on frames captured from CCTV cameras or webcam sources.

A. System Pipeline Flowchart

The overall workflow of the NIRBHAY system is illustrated in Fig. 1. Each video frame undergoes preprocessing before being independently processed by the three detection modules. The results from all modules are then rendered together on the application interface.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

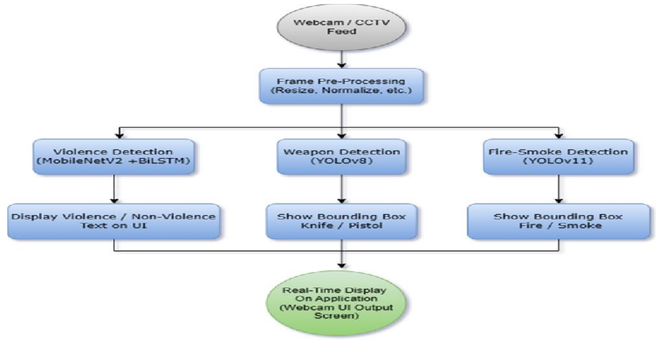


Fig. 1: NIRBHAY System Pipeline Flowchart

The step-by-step pipeline is described below:

- Camera Input: Live video is continuously captured from a webcam or CCTV feed.
- Frame Pre-processing: Each frame is resized, normalized, and formatted according to model requirements.
- Parallel Threat Detection: Three modules operate simultaneously -Violence Detection (MobileNetV2 + BiLSTM), Weapon Detection (YOLOv8), and Fire & Smoke Detection (YOLOv11).
- User Interface Output: Detection results are displayed in real time, with text labels for violence detection and bounding boxes for detected weapons, fire, and smoke.
- Continuous Monitoring: This process repeats for every incoming frame, ensuring uninterrupted real-time surveillance.

B. Dataset Description

The system was trained using three publicly available benchmark datasets, each corresponding to a specific detection module. The details are summarized in Table I.

TABLE I: Dataset Description

Module	Dataset	Classes	Size	Annotation
Violence Detection	Real-Life Violence Situations Dataset	Violence, Non-Violence	2000+ videos	Frame sequences
Weapon Detection	Guns and Knives Detection Dataset	Pistol, Knife	8000+ images	YOLO BBox labels
Fire-Smoke Detection	D-Fire Dataset	Fire, Smoke	6000+ images	YOLO BBox labels

C. Data Pre-processing

Different preprocessing strategies were applied for each module based on its requirements.

Violence Module (MobileNetV2 + BiLSTM):

Video data was divided into short temporal clips by extracting 16 consecutive frames per sample. Each frame was resized to 96×96 pixels and normalized to [0,1]. The dataset was balanced between violence and non-violence classes to reduce bias during training.

Weapon Module (YOLOv8):

All images were verified to ensure a correctly formatted YOLO label file. Images were resized to 640×640 as per YOLOv8 requirements. Proper train, validation, and test splits were maintained to ensure reliable evaluation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Fire-Smoke Module (YOLOv11):
Annotations were validated to match YOLOv11 standards. Additionally, a class distribution analysis was performed across fire-only, smoke-only, and mixed hazardous environment scenes to identify and address dataset imbalances.

D. Feature Extraction
Each module uses specialized deep learning architectures to extract relevant features from the input data.

Violence Detection:
MobileNetV2, pretrained on ImageNet, is applied to each frame using Time Distributed layers to extract spatial features. These features are then passed to a Bidirectional LSTM, which captures temporal patterns and motion dynamics across frame sequences.

Weapon Detection:
YOLOv8 employs a CSP-based backbone to extract high-level spatial features efficiently. Its PAN-FPN architecture enables detection of objects at multiple scales, which is important for identifying small or partially occluded weapons.

Fire-Smoke Detection:
YOLOv11 utilizes an enhanced CSP backbone to improve detection in challenging conditions such as low visibility or diffuse smoke patterns.

E. Model Architectures
E.1 Violence Detection Model (MobileNetV2 + BiLSTM)
The violence detection module combines spatial and temporal learning: MobileNetV2 for spatial feature extraction per frame, and a Bidirectional LSTM for temporal motion modeling across 16-frame sequences. The final output is a binary classification indicating Violence / Non-Violence.

TABLE II: Violence Model Configuration

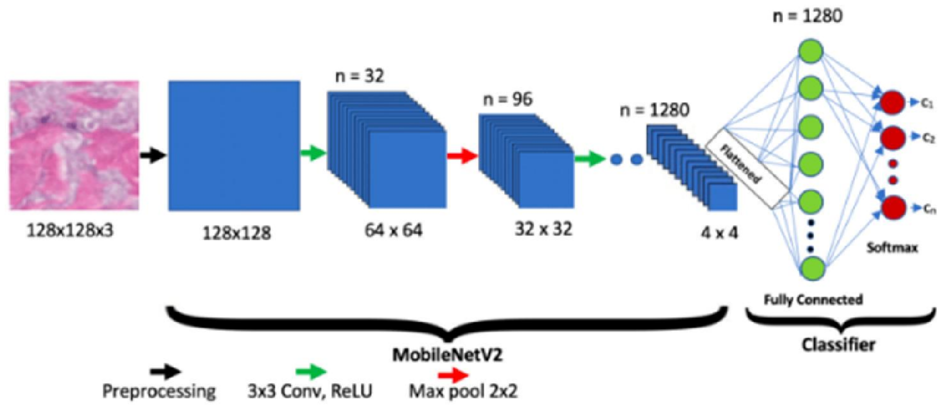
Component	Purpose
MobileNetV2 (ImageNet)	Spatial feature extraction per frame
TimeDistributed Wrapper	Applies CNN independently to each frame
BiLSTM	Temporal motion learning (forward + backward)
BatchNorm + Dropout	Regularization
Dense + Softmax	Binary classification: Violence / Non-Violence

TABLE III: Violence Model Hyperparameters

Parameter	Value
Input Shape	(16, 96, 96, 3)
Output Classes	2 (Violence / non-violence)
Optimizer	SGD
Regularization	L2 + Dropout
Epochs	50

E.1.1 MobileNetV2 Architecture

MobileNetV2 is highly efficient and lightweight, making it ideal for real-time video analysis. Its inverted residual and depth wise separable convolution layers extract strong spatial features from each frame while keeping computation and model size low. Fig. 2 illustrates the feature extraction flow.



E.1.2 BiLSTM Architecture

Violence involves rapid and complex motion patterns that evolve over time. A Bidirectional LSTM learns temporal dependencies in both forward and backward directions, enabling it to understand sudden movements or aggressive gestures across frames. Fig. 3 shows the BiLSTM unrolled across time steps.

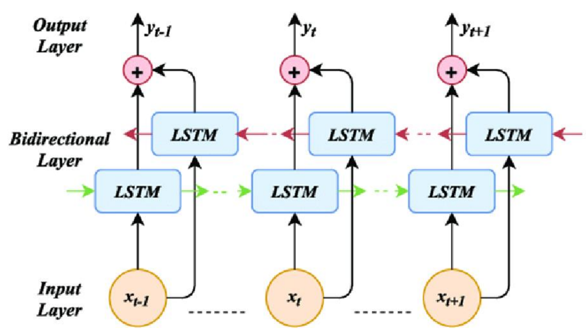


Fig. 3: Bidirectional LSTM (BiLSTM) — Forward and Backward Temporal Modeling

E.2 Weapon Detection Model (YOLOv8)

The weapon detection module uses YOLOv8 as the primary feature extractor and object detector. Its CSP-based backbone with PAN/FPN neck supports multiscale detection, critical for identifying small and partially occluded weapons such as pistols and knives in CCTV footage. Fig. 4 shows the YOLOv8 architecture.

TABLE IV: YOLOv8 (Weapon) Configuration

Parameter	Value
Input Resolution	640×640
Loss	Cls + Obj + Box
Optimizer	SGD

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Epochs	40
Inference Speed	>30 FPS (Real-time)

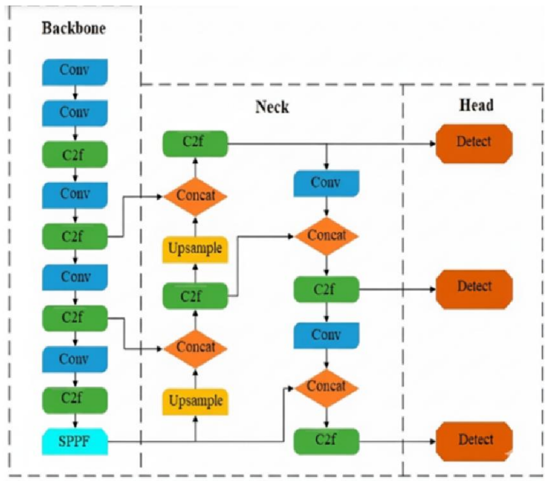


Fig. 4: YOLOv8 Architecture — Backbone, Neck (PAN-FPN), Detection Head, and Soft-NMS

E.3 Fire-Smoke Detection Model (YOLOv11)

Fire and smoke have irregular shapes and often blend into backgrounds, requiring deeper feature extraction and stronger generalization. YOLOv11's enhanced CSP Backbone v2 and improved detection head deliver robust identification of fire-smoke regions under real-world conditions. Fig. 5 illustrates the YOLOv11 architecture.

TABLE V: YOLOv11 (Fire-Smoke) Configuration

Parameter	Value
Input Resolution	640×640
Epochs	100
Batch Size	8
Optimizer	SGD
Regularization	Weight Decay + Patience (Early Stopping)

DOI: 10.48175/IJARSCT-36880

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

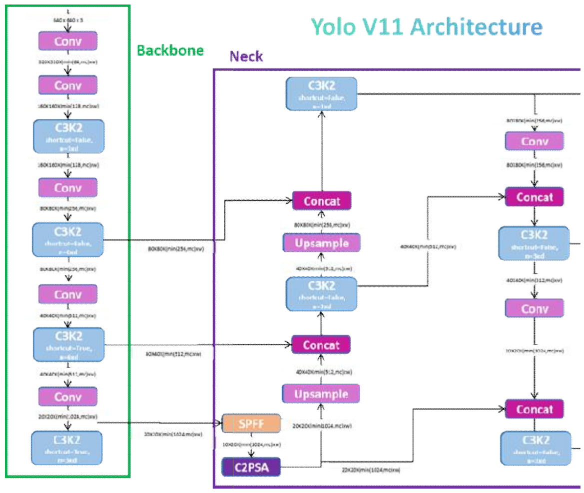


Fig. 5: YOLOv11 Architecture — Enhanced CSP v2 Backbone, FPN+PAN Neck, Soft-NMS Head

F. Tools and Techniques Used

Category	Tools / Frameworks
Programming Language	Python
Deep Learning Libraries	TensorFlow, Ultralytics YOLO
Computer Vision	OpenCV
Development Environment	Kaggle Notebook, VS Code
GPU Support	NVIDIA Tesla P100 (Kaggle GPU Runtime)

IV. RESULTS AND DISCUSSION

The NIRBHAY system was trained and evaluated across three independent detection modules. Each module was assessed using standard performance metrics including Precision, Recall, F1-Score, and mean Average Precision (mAP). The results confirm that the integrated pipeline achieves reliable real-time threat detection across all three threat categories.

A. Violence Detection Module (MobileNetV2 + BiLSTM)

The violence detection module was trained for 50 epochs on the Real-Life Violence Situations Dataset. Both training and validation loss curves exhibited a smooth and consistent decline, converging from an initial value of approximately 14.5 to near 0.5 by epoch 50. Simultaneously, training and validation accuracy rose rapidly from approximately 60% in the first few epochs, stabilising above 95% from epoch 15 onwards and approaching 99% by the final epoch, with validation accuracy closely tracking the training curve — indicating stable learning without overfitting. Fig. 1 shows both curves side by side.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

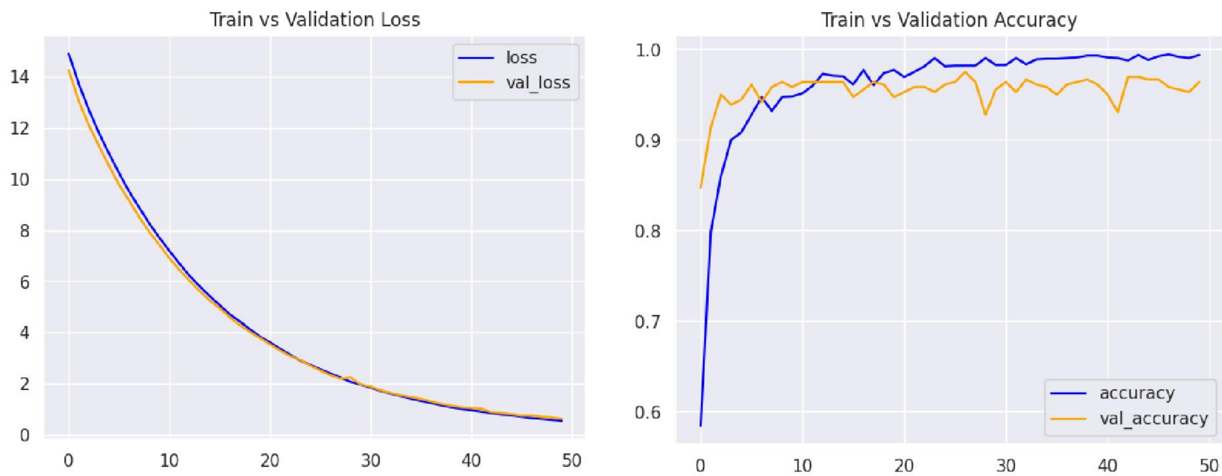


Fig. 1. Train vs. Validation Loss (left) and Train vs. Validation Accuracy (right) — Violence Detection Module

The confusion matrix, evaluated on a held-out test set of 200 samples (99 non-violence, 101 Violence), shows 97 correct non-violence predictions and 97 correct Violence predictions, with only 2 false positives and 4 false negatives. Fig. 2 presents the confusion matrix, and Table I provides the full classification report.

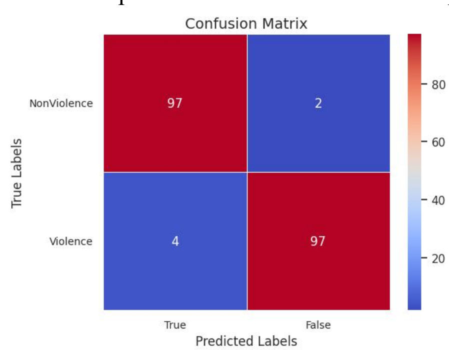


Fig. 2. Confusion Matrix — Violence Detection Module

TABLE I: Violence Detection — Classification Report

Class	Precision	Recall	F1-Score	Support
0 — Non-Violence	0.96	0.98	0.97	99
1 — Violence	0.98	0.96	0.97	101
Accuracy	—	—	0.97	200
Macro Avg	0.97	0.97	0.97	200
Weighted Avg	0.97	0.97	0.97	200

The module achieves an overall accuracy of 97% with a macro-averaged F1-Score of 0.97, demonstrating balanced and high performance across both classes. The near-equal precision and recall for both Violence (0.98 / 0.96) and non-

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

violence (0.96 / 0.98) confirm that the MobileNetV2 + BiLSTM architecture effectively captures both spatial appearance features and temporal motion cues required for reliable violence classification.

B. Weapon Detection Module (YOLOv8)

The weapon detection module was trained using YOLOv8 on the Guns and Knives Detection Dataset over 40 epochs, with the best model checkpoint recorded at epoch 34. All three loss components — bounding box loss, classification loss, and distribution focal loss — showed consistent decreasing trends across both training and validation sets. Fig. 4 shows the full training metrics grid including all loss curves, precision, recall, and mAP progression.

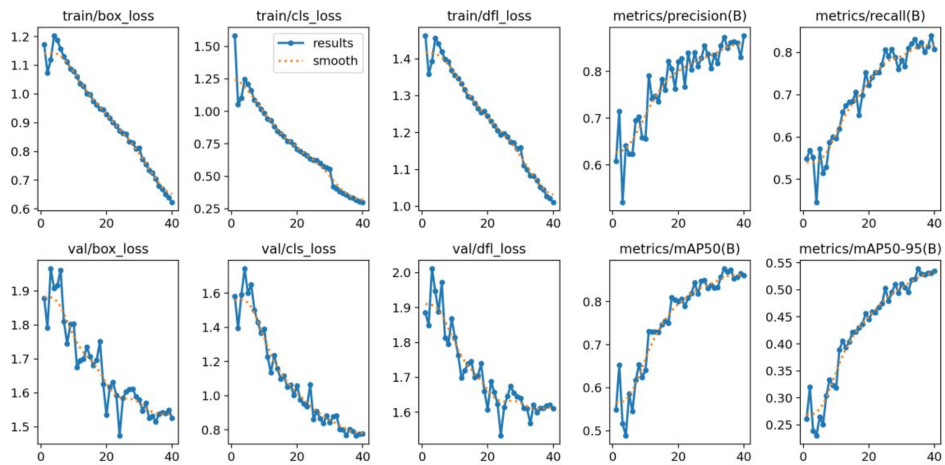


Fig. 4. Training and Validation Loss, Precision, Recall, and mAP Curves — Weapon Detection Module

The confusion matrix shows 168 correct pistol detections and 109 correct knife detections, with 33 pistol false negatives and 19 knife false negatives. Background false positives numbered 54 for pistol and 8 for knife, indicating high specificity especially for the knife class. Fig. 5 presents the confusion matrix, and Table II summarises the final performance metrics at the best checkpoint.

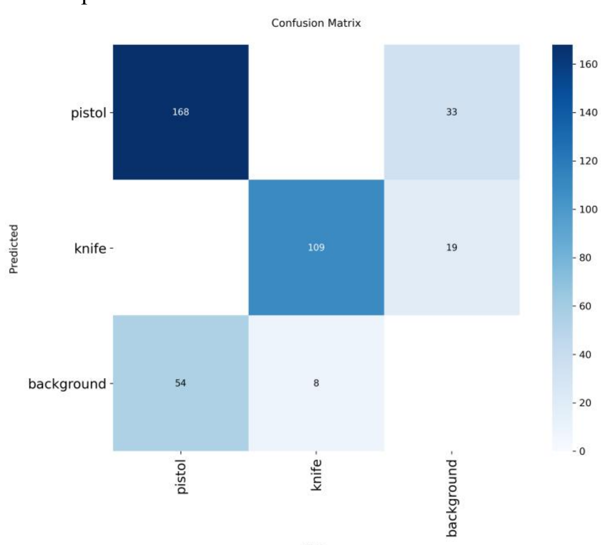


Fig. 5. Confusion Matrix — Weapon Detection Module

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

TABLE II: Weapon Detection Module — Performance Metrics (YOLOv8, Best at Epoch 34)

Metric	Value
Epoch (Best Checkpoint)	34
Precision	0.8733
Recall	0.8307
mAP@50	0.8764
mAP@50-95	0.5200

The weapon module achieves a mAP@50 of 0.8764, reflecting strong localisation and classification capability for both pistol and knife classes in CCTV-style imagery at real-time inference speeds exceeding 30 FPS. The mAP@50-95 of 0.520 demonstrates robustness across varying IoU thresholds. The slight gap between precision (0.873) and recall (0.831) suggests that a small proportion of weapon instances — particularly at distance or under partial occlusion — are occasionally missed, consistent with known small-object detection challenges in surveillance footage.

C. Fire-Smoke Detection Module (YOLOv11)

The fire-smoke detection module was trained using YOLOv11 over 100 epochs on the D-Fire dataset, covering fire, smoke, and background classes. All loss curves for both training and validation sets exhibited consistent convergence throughout the full training run. Fig. 6 shows the complete training metrics grid including loss curves, precision, recall, and mAP progression over 100 epochs.

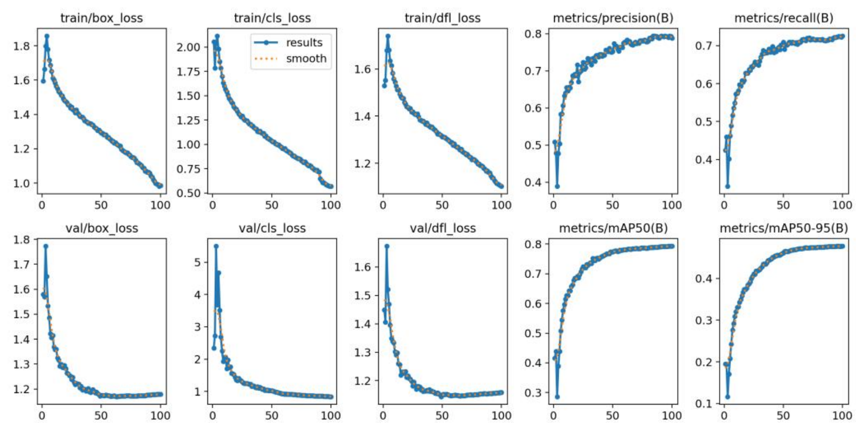


Fig. 6. Training and Validation Loss, Precision, Recall, and mAP Curves — Fire-Smoke Detection Module

The confusion matrix shows 1482 correct smoke detections and 1601 correct fire detections. Moderate background misclassifications were observed — 268 smoke and 559 fire instances classified as background — primarily attributable to low-contrast or transparent early-stage smoke regions. Fig. 7 presents the confusion matrix, and Table III summarises the final evaluation metrics.

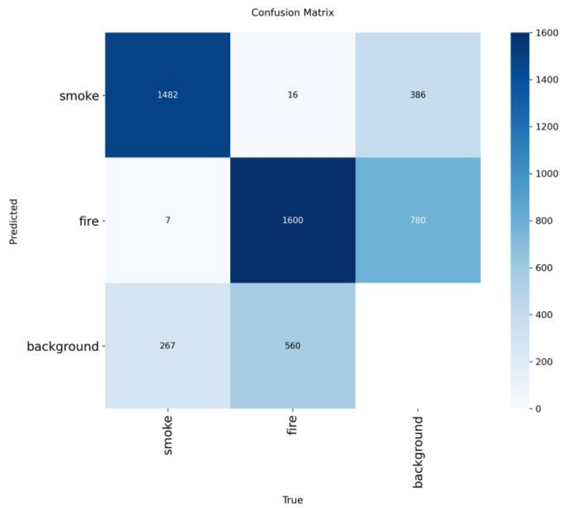


Fig. 7. Confusion Matrix — Fire-Smoke Detection Module

TABLE III: Fire-Smoke Detection Module — Performance Metrics (YOLOv11, Epoch 100)

Metric	Value
Epoch (Best Checkpoint)	100
Precision	0.7880
Recall	0.7261
mAP@50	0.7932
mAP@50-95	0.4775

The fire-smoke module achieves a mAP@50 of 0.7932, demonstrating reliable fire and smoke localisation across varied indoor and outdoor environments. The lower recall (0.726) relative to precision (0.788) is primarily attributable to missed detections in challenging scenarios such as transparent early-stage smoke that blends into the background. The mAP@50-95 of 0.4775 reflects moderate bounding box precision for diffuse hazard regions and points toward an area for future improvement.

D. Comparative Summary Across All Modules

Table IV presents a consolidated performance comparison of all three NIRBHAY detection modules.

TABLE IV: Comparative Performance Summary — All NIRBHAY Modules

Module	Architecture	Precision	Recall	mAP@50 / F1	Best Epoch
Violence Detection	MobileNetV2 + BiLSTM	0.97	0.97	0.97 (F1)	50 / 50
Weapon Detection	YOLOv8	0.8733	0.8307	0.8764	34 / 40
Fire-Smoke Detection	YOLOv11	0.7880	0.7261	0.7932	100 / 100

The comparative summary confirms that NIRBHAY delivers strong performance across all three threat categories. The violence detection module achieves the highest balanced metrics (F1-Score = 0.97), benefiting from the BiLSTM's ability to model temporal motion dynamics across 16-frame sequences. The weapon detection module delivers high localisation accuracy (mAP@50 = 0.876) at real-time inference speeds exceeding 30 FPS. The fire-smoke module achieves a competitive mAP@50 of 0.793, validating the YOLOv11 enhanced CSP backbone for diffuse hazard detection. Together, the three modules form a comprehensive and integrated multi-threat real-time surveillance system.

V. CONCLUSION

This paper presented NIRBHAY, an AI-powered multimodal threat detection system integrating MobileNetV2+BiLSTM for violence detection, YOLOv8 for weapon recognition, and YOLOv11 for fire and smoke detection into a unified real-time surveillance framework.

Experimental results confirm strong performance across all modules — a violence detection F1-Score of 0.97, weapon detection mAP@50 of 0.8764 at over 30 FPS, and fire-smoke detection mAP@50 of 0.7932 — demonstrating that the system is both accurate and practically deployable.

NIRBHAY significantly reduces dependency on continuous human monitoring and improves incident response time. Its modular design allows individual components to be independently updated as better models and datasets become available.

Future work includes incorporating audio analysis, adopting Vision Transformers, expanding weapon categories, optimizing for edge devices, and integrating automated emergency alerts for real-world deployment.

VI. ACKNOWLEDGMENTS

"We would like to express our sincere gratitude to the Department of Computer Science and Department of Data Science, Fergusson College, Pune, for providing the necessary resources and support throughout this research."

REFERENCES

[1] A. L. E. Reyes and L. Gomez, "Anomalous weapon detection for armed robbery using YOLOv8," *Applied Sciences*, vol. 15, no. 4, p. 3452, 2025. doi: 10.3390/app15043452.

[2] A. B. S. Mahi and N. Dündar, "A comprehensive dataset for violence detection in surveillance," *Data in Brief*, vol. 58, p. 109247, 2024. doi: 10.1016/j.dib.2024.109247.

[3] D. Berardini, M. Ricci, and P. Romano, "Edge AI and super-resolution for resource-constrained weapon detection," *Expert Systems with Applications*, vol. 242, p. 122765, 2025. doi: 10.1016/j.eswa.2025.122765.

[4] E. B. Varghese and K. Menon, "A temporal–spatial attention framework for fine-grained violence detection," *Multimedia Tools and Applications*, vol. 84, no. 3, pp. 10123–10138, 2025. doi: 10.1007/s11042-025-19654-4.

[5] I. A. Haiura, "Detecting violence in videos using 3D CNNs," *Pattern Recognition Letters*, vol. 178, pp. 1–10, 2024. doi: 10.1016/j.patrec.2024.06.004.

[6] K. Akhila and K. R. Ahmed, "Real-time deep learning weapon detection techniques for mitigating lone wolf attacks," *arXiv preprint arXiv:2405.14148*, 2024.

[7] M. Chetoui, C. Bahhar, and M. Benabdellah, "Fire and smoke detection using fine-tuned YOLOv8 and comparative study," *Sensors*, vol. 24, no. 2, p. 950, 2024. doi: 10.3390/s24020950.

[8] M. Talib and S. Khan, "Multi-weapon detection using deep learning approaches," *Int. J. Inf. Commun. Technol.*, vol. 18, no. 2, pp. 45–54, 2024.

[9] P. Negre, J. Gutiérrez, and S. Kim, "Literature review of deep-learning-based video violence detection," *Sensors*, vol. 24, no. 3, p. 1154, 2024. doi: 10.3390/s24031154.

[10] R. Dugyala and P. Babu, "Weapon detection in surveillance videos using YOLOv8," *E3S Web Conf.*, vol. 450, p. 01023, 2023. doi: 10.1051/e3sconf/202345001023.

ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

[11] R. Vijeikis, R. Mikalauskas, and E. Paulauskas, "Efficient violence detection in surveillance videos using lightweight CNN models," *Electronics*, vol. 11, no. 12, p. 1921, 2022. doi: 10.3390/electronics11121921.
[12] V. E. Sathishkumar and M. Ramesh, "Forest fire and smoke detection using deep learning," *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, no. 1, p. 25, 2023. doi: 10.1186/s42492-023-00112-6.
[13] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 206–221.

DOI: 10.48175/IJARSCT-36880