An Efficient Detection of Suspicious Objects from Dynamic Video Surveillance by Fusion-based Multiview Deep Learning Techniques

Ramesh Chandra Poonia[1], Kamal Upreti[1,*], Nidhi Singh[2], Jyoti Kesarwani[3] and Mohammad Shabbir Alam[4]

[1]*Department of Computer Science, Christ University, Delhi NCR, India*
[2]*G. D. Goenka University, Gurgaon, Haryana, India*
[3]*United College of Engineering and Research, Prayagraj, Uttar Pradesh, India*
[4]*Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, KSA*
E-mail: rameshcpoonia@gmail.com; kamalupreti1989@gmail.com; nidhi.singh@gdgu.org; jyotiecimt@gmail.com; amushabbir@gmail.com
[*]*Corresponding Author*

Received 10 October 2024; Accepted 15 December 2024

Abstract

Real-time detections of suspicious objects are needed to identify for finding criminal activities and are used in immediate alert systems for public safety applications. Video surveillance systems use live, closed-circuit televisions (live CCTVs) for dynamic video capturing of objects. Finding criminal activities over the dynamic video data is an emerging surveillance problem. The deep learning techniques are tedious for detecting suspicious movable objects and criminal activities. YOLO (You Only Look Once) gives more prominent movable video object detection accuracy than conventional deep models, like Convolutional Neural Network (CNN), 3D CNN, and Convolutional LSTM. State-of-the-art YOLO models, YOLOv8n, YOLOv8s, and YOLOv8l, are emphasized for extracting and detecting object motion detection from the

Journal of Mobile Multimedia, Vol. 21_1, 1–26.
doi: 10.13052/jmm1550-4646.2111

dynamic video. YOLO models use single-view deep learning to classify or detect objects. These models limit the accuracy of the detection of complex and dynamic objects of dynamic video data. This paper presents the Fusion-based Multiview deep learning techniques to overcome this issue. The experimental study demonstrates that the proposed methodology efficiently detects suspicious data objects more than the single-view deep models.

Keywords: Deep learning, video surveillance, object detection, YOLO and fusion.

1 Introduction

Video surveillance [1] is one of the most significant social applications for identifying criminal activities and detecting suspicious objects in public places. Video surveillance and visual learning are helpful in shopping malls for the dynamic detection of crimes and in traffic for the detection of suspected vehicles. Notably, the deep models, CNN, 3D CNN, ConvLSTM [2–4] are the recommended techniques for object classification in computer vision. YOLO [5] is the most advanced deep video model for dynamic object detection and classification. In artificial intelligence (AI) implementations [6], dynamic learning with YOLO models has enabled effective dynamic scene interpretations. Three YOLO variants [7–9] ((YOLOv8n, YOLOv8s, YOLOv8l) have been developed recently for the deep learning of dynamic video object detection. These models give impressive computer vision classification results and may be helpful for the automation of AI-based visual systems for the instance detection of crime activities. Computer vision is an emerging field of AI in which deep learning and machine learning techniques play immense roles in digital classifications. There is a wide range of social applications of computer vision, including public security systems, trajectory predicted systems (TPS) [10] for suspicious vehicles, healthcare systems, industries, etc. The YOLO models are more prominent in visual systems for tracking and suspecting objects based on trained shapes from dynamic video data. The YOLO takes only the forward pass in the network. It indicates that the 'only look once' mechanism is implemented. All the diversified objects are well trained using the efficient architecture of the YOLO model [11] (shown in Figure 1). The architecture of YOLO was specially designed to classify movable objects and may become most valuable in real-time video surveillance applications. The YOLO architecture consists of a potent Darknet framework [12], while other deep models do

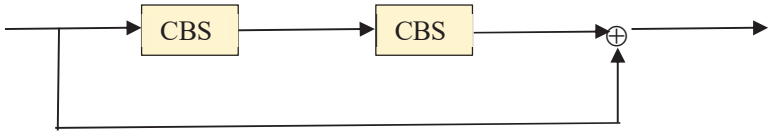
not have this framework. This DarkNet aims to extract the abstract features of movable objects by designing consecutive constructions of convolution layers. Initially, the dynamic frames or images of the scene are fed with the pixel size of 448x448x3; i.e., it takes the input frame and resizes with 448*448, and also padding operations are performed to maintain the same aspect ratio. Next, resized frames are forwarded to the next level of YOLO architecture, DarkNet. The DarkNet framework processes the frames with a sequence of convolution and max-pooling layers. The output is flattened with the size of 4096 and produced the size of a 7x7 grid by the fully connected layers. Significant research has progressed in the YOLOv8 deep model for the detection of movable (or dynamic objects), which are as follows: extensive ideas of small object detection, enhanced backbone model of YOLO, and focus on optimizing the loss function. The methodology of bi-directional feature pyramid networks (Bi-PAN-FPN) [13] is used in the extension of YOLOv8, and its enhanced YOLOv8 model improves the detection capacity of small objects. Another concept used in YOLOv8 is integrated Ghostblock Unit and Wise Intersection over Union (WiseIoU) [14] bounding regression loss. Its integrated technique is mainly used to optimize the loss of the model in small object detections. Figure 2 shows the bounding box of the YOLO model for detecting objects in a sample frame of the dynamic video. Key operations of the YOLOv8 are the extraction of features, enhancing the features, and the detection of the object's predictions. Totally, it's architecture consists of three major components, namely, backbone, neck, and head. The backbone that extracts features, the neck to refine the features, and the head that predicts object labels and bounding boxes. These three components work in tandem to enable efficient detection of objects in dynamic video data with high accuracy and speed. With these components it can be efficiently detects the objects from the dynamic video. The input frames of the images are initially processed by the primary component of YOLOv8, i.e. backbone. It uses the 3x3 convolution layers using the stride size of two for extraction of features from the either frames or images of video data. The cross-stage partial (CSP) is the crucial part of YOLOv8 which has obtained from [15] and replaces the three successive convolutions with two convolutions in its pre-processing stage. It shown the greater improvement in improving the rate of model's speed by reducing the number of convolutions. Architecturally, this modification precisely aligns with the purposes of this work toward optimized detection performance without affecting the accuracy of detection. Second component of YOLOv8 is neck, which used for refinement of extracted features of earlier backbone component and the refined features for the next

stage of detection. Final head playing the crucial role to predict the objects labels and detected object' s bounding boxes. Key terminologies of YOLOv8 model are as follows:

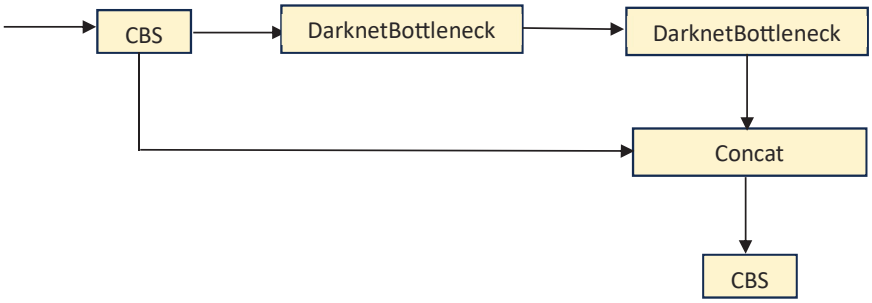
ConvModule: This module presents the convolution layers and then it uses the batch normalization with the Sigmoid Linear Unit (SiLU) activation function. This function more smoothens the features than the ReLU function.

CBS = → Convolution → Batch Normalization → SiLU

Bottleneck: The bottleneck layers in the YOLOv8 architecture poses the input data in the form of lower dimensional representation. Thus, complexity of computations are reduced with the layers of bottleneck.



C2f: It consists of two convolutions and bottleneck block that reduces the number of channels of the features



Spatial pyramid pooling fusion (SPPF): The SPPF layer is very important to increase the training speed of the model. It consists of the CBS and number of max-pooling layers. Each and every result of max-pooling layer is further concatenated and the final concatenated features are fed into the convolutional block. Figure 2 shows the implementation of bounding box stages in YOLO. Initially, the frame (or image) of dynamic video should fixed in the grid. Grid cells are assigned for the object's identification. The bounding box in the image specifies the spatial information of the objects. It is rectangular in shape with an upper-left corner and a lower-right corner. YOLO efficiently performs the prediction of bounding boxes toward

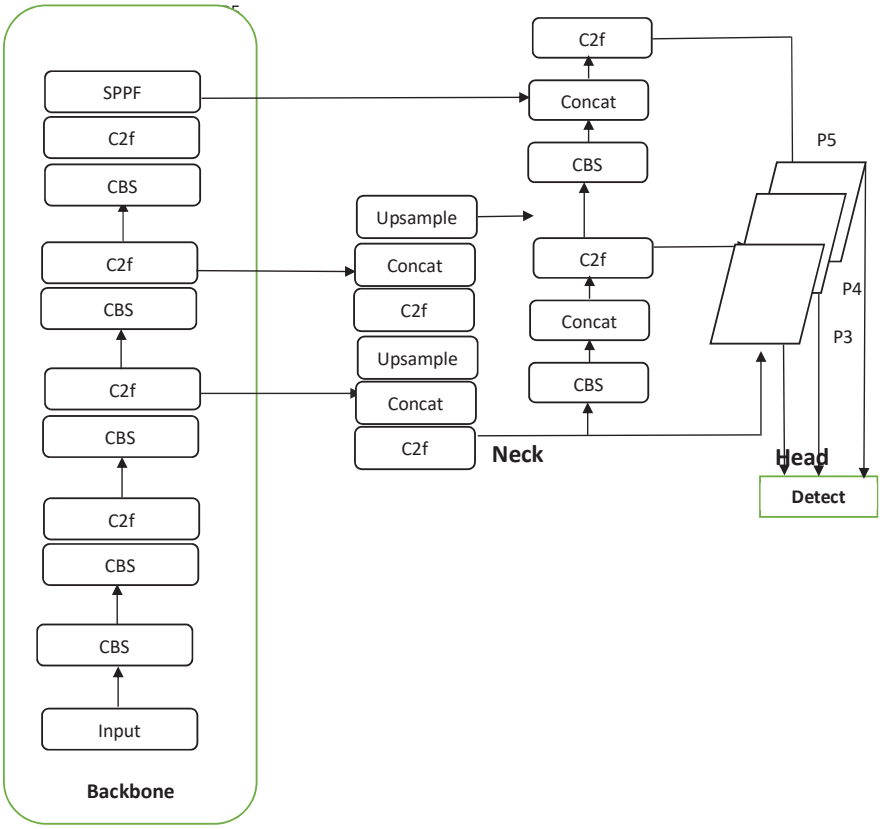
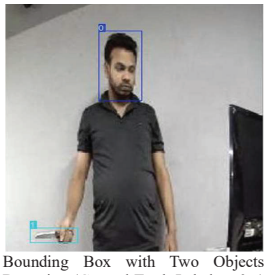
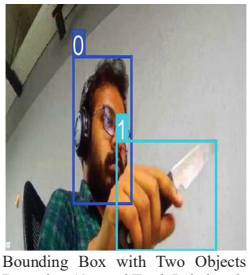


Figure 1 YOLOv8 architecture with convolution, flattened, and other layers.



i. Bounding Box with Two Objects Detection (Ground Truth Labels – 0, 1 for Face, Knife Detection)



ii. Bounding Box with Two Objects Detection (Ground Truth Labels – 0, 1 for Face, Knife Detection)

iii. Bounding Box with Two Objects Detection (Ground Truth Labels – 0,1, and 2 for Face, Knife, and Pistol Detection)

Figure 2 Bonding boxes for object detection in YOLO models.

the objects' pinpointing locations. State-of-the-art YOLOv8 model architectures, YOLOv8n, YOLOv8s, and YOLOv8l, support the faster detection of movable objects and are capable of detecting suspicious objects instantly. Computer vision is underlying the rapid growth of visual transformations; it has significantly played a vital role in visual social applications, including the identification of fraudulent or criminal activities by detecting suspected video objects, suspected vehicles in the traffic, detection irregularities in the EHRs for healthcare [15], etc. YOLO is an efficient deep visual model for learning and classification tasks in all such cases. Three variants of YOLO, nano, small, and large, were developed for efficient object recognition tasks, pronounced as YOLOv8n, YOLOv8s, and YOLOv8l, respectively. These methods follow the single view deep visual learning. More than learning with a single deeper model may be required for complete communication or detection tasks. Multiview deep models can learn (or train) more effectively with multiple fusion models for dynamic and complex shaped objects. Thus, fusion-based multiview deep learning techniques are proposed in this paper for the detection of such dynamic objects detection as per the needs of video surveillance applications.

The way of organization of the paper is mentioned as follows: Study of the deep models for video object detections presented in Section 2. Proposed fusion-based multiview deep learning techniques are neatly described and discussed in Section 3. Further, the results analysis and its comparative study are discussed in Section 4. Finally, the conclusion is described in Section 5.

2 Study of the Deep Models for Video Object Detections

Nowadays, deep models play a crucial role in visual detection and recognition tasks compared to other machine learning models. The current era of AI uses deep models to develop automation tasks rapidly. With the deep models, CNN, 3D CNN, LSTM [16], and ConvLSTM [17] achieved a good classification rate with optimized learning for text, image, and audio conventional applications. However, video mining advances need to classify dynamic video objects. State-of-the-art deep models [18–20], YOLO, and its extension models are capable of retrieving the features of dynamic or movable objects of video surveillance data. These are useful, for instance, identification of movable objects by the efficient YOLO deep learning. This literature section presents the study of video deep models. Initially, the YOLOv1 was developed, and the architecture consists of a sequence of a large number of convolutional layers, flattened, and fully connected layers. Consequently,

other variants of YOLO models, YOLOv2.... YOLOv8 has been developed recently. YOLOv1 [21] Posed the grid cell with the size of m x m onto a frame or image and centered on the object's location for detection. The imposed grid cell makes the bounding box with the two parameters of confidence score and dimensions. The probability value for the object in the specified bounding box is derived, which is helpful in computing the value of the confidence score. Equation (1) denotes the computing formula of confidence score.

$$\text{Confidence score} = \text{p}(\text{object}) \times \text{IoU}_{\text{tp}} \tag{1}$$

The variable 'p' finds the probability of the presence of an object (the value is between 0 and 1) in the bounding box, whereas another variable, 'IoU_{tp}', takes intersection over union for the predicted bounding box and the ground truths [22]. YOLO aims to detect and localize objects within the bounding boxes accurately. Challenges of overlapping bounding boxes addressed by YOLO with the mechanism of non-maximum suppression (NMS). With this mechanism, some are eliminated in which the IoU value is below the target threshold value. The advantage of NMS is that detects the object with a higher value of IoU; i.e., the object with a higher value of IoU with the predicted and ground bounding boxes indicates that the object detected is more accurate. The IoU is one of the best metrics for object detection in computer vision. YOLOv1 has two limitations: recall problems and localization errors when detecting the objects. YOLOv2 is another subsequent enhanced model that handles these issues by its notable design. This model includes the Darknet-19 framework with 19 convolution layers and five maximum pooling layers [23]. It uses various pooling and 1×1 convolution to find the notable downsampling in the architecture. YOLOv2 models tackle the issue of the limited availability of video-labeled data by training the ImangeNet and COCO datasets. It improves the recall rate by 7% when compared to the YOLOv1 model for object detection since it uses the 19 convolutional layers for the feature extraction and bounding boxes prediction done by the anchor boxes. The localization error problem remains the same in YOLOv2. YOLOv3 [24] is another successive model whose architecture has 53 convolutional layers and several residual connections. The key component of convolutional layers and residual connection in the YOLOv3 is referred to as the Darknet-53. In Darknet-53, it changes all the maximum pooling layers, places the stride convolutions, and combines with residual layers. It reduces the localization loss by associating the anchor boxes with every ground truth object. The problem of YOLOv3 is that it depends on a single anchor point for detecting the ground truth object. There

is a chance of getting a lower classification rate in some real-time video applications. To overcome this problem, YOLOv4 [25] developed with the integration of CSP Darknet53, spatial pyramid pooling (SPP) [26], PANet structure [27], cross iteration batch normalization network architecture [28], and segment anything model (SAM) [29] incorporation. YOLOv4 uses multiple anchors and their efficient architecture to associate ground truth objects. There is a chance to get more positive anchors in the selection, unlike a single selection in YOLOv3, to improve bounding detection accuracy. It shows a significant improvement in localization accuracy rate by maximizing the IoU and regularization. YOLOv5 [30] is another developed model that uses PyTorch rather than constructing the framework of Darknet. It derives the features of a cross-stage partial net from the ResNet. The head part of the YOLOv5 consists of the convolutional layers for bounding box predictions and labels. It improved feature extraction and aggregation, as well as anchor-based predictions. A higher average precision performance was achieved in YOLOv5 compared to YOLOv1 to YOLOv4 for the high quality of images. Yolov5 differs from the predecessors because it uses PyTorch instead of Darknet, which was traditionally used in the previous versions of YOLO. PyTorch is better in terms of flexibility, scalability, and usability for model development and deployment. YOLOv5 incorporates the idea of Cross-Stage Partial Network (CSPNet) inspired by ResNet for better feature extraction and model efficiency. CSPNet splits the feature map into two segments to reduce the computational cost as well as enhance gradient flow, which makes YOLOv5 better on accuracy and speed with lower resources. This makes YOLOv5 quite efficient for a variety of object detection tasks while harnessing the robust training abilities of PyTorch. YOLOv6 [31] was developed for industry applications for single-stage object detection. It refines the architectures by augmenting CSPDarknet. It increases the scale of the features, which helps to improve the detection accuracy compared to YOLOv5. The detection process processes an average of 50 frames per second (FPS) based on the processor requirement of the T4 GPU. To improve the scalability, YOLOv7 [32] was designed with an extended efficient layer aggregation network (E-LAN). This model performs efficient and scalable learning and increases the FPS by 150. The key benefits of the YOLOv7 model are accuracy detection and scalability improvements. The YOLOv8 [33] model is currently used for effective video-learning applications. A distinct feature of YOLOv8 is to adopt the anchor-free bounding boxes by the object's center predictions. Small objects are also effectively tackled and detected, so its research has greatly succeeded in a wide range of AI tasks of computer

vision: object detection, frame segmentation, and dynamic classification of videos. In YOLOv8, four variants were designed with model sizes of n, s, m, and l (referred to as nano, small, medium, and large, respectively). These four models are referred to as YOLOv8n, YOLOv8s, and YOLOv8l, [34] which are derived based on the size and depth of the network architecture. YOLOv8n is a light model with fewer layers and less depth architecture. It finds the detection results in a faster manner; however, it is required to produce more potent and accurate detection results. This is the reason for modifying the architecture sizes in terms of the width and depth of the layer, which finally imposes an increase in the complexity of architecture in the models from YOLOv8s to YOLOv8l. Variants of YOLOv8 models are recommended based on the object complexity, processor availability, and other scalability estimations. Balancing speed and detection accuracy are the most emerging aspects of object classification and recognition tasks. Live CCTV is one important method of detecting suspicious objects for real-time identification of potential crimes, allowing an intervention to be made swiftly and enhancing public safety. The state-of-the-art YOLO models especially cater to dynamic video data: high accuracy in object classification and detection. Their advanced architectures comprise YOLOv8 variants, ensuring reliable and efficient performance in high-quality video surveillance systems, which render them indispensable tools for preventing crime and public security.

3 Fusion-based Multiview Deep Learning Techniques

Two fusion techniques were developed for the implementation of Multiview deep learning. These fusion techniques have been developed based on the strengths of YOLOv8-nano and YOLOv8-large models. Weight-level fusion involves averaging the trained weights of both models to develop a new fused model, with the speed of YOLOv8-nano combined with the accuracy of YOLOv8-large. Feature-level fusion integrates features extracted from both models to capture a broader range of object characteristics, thereby enhancing detection accuracy for dynamic and complex scenes. Weight and feature compatibility across the development process ensured optimization of performance without architectural complexity increase. Weight-level fusion is lightweight, which balances speed and efficiency, making it very suitable for real-time applications with limited computational resources. Feature-level fusion is meant to achieve higher accuracy; hence, it is more appropriate for detecting complex and diverse objects in dynamic scenarios.

Table 1 Comparative analysis of YOLOv8 variants

Variant	Architecture Complexity	Speed	Accuracy	Best Use Case
YOLOv8n	Lightweight	Fast	Moderate	Real-time, resource-constrained applications
YOLOv8s	Balanced	Moderate	High	Applications needing both speed and accuracy
YOLOv8l	Complex	Slower	Very High	High-resolution or complex object detection

Single-view deep learning for video object detection currently uses the YOLO models. YOLO models rely on single-view deep learning; this is because they are only able to process individual frames, or views, of the video data at any particular time, and the use of fast and efficient object detection. The YOLO models have the core component of Darknet architecture and flattened and fully connected layers. These models improve the detection accuracy compared to traditional deep models. The YOLOv8 is currently being successively used the technique on video surveillance applications. The YOLO models generate the bounding boxes to detect objects. The parameter of IoU is taken as a critical evaluation measure for predicting the bounding measure based on ground truth bounding measures. Three distinct models of YOLOv8 are nano, small, and large, which are distinguished mainly by the complexity of architecture as depicted in Table 1. The nano model has a less complex architecture (i.e., fewer layers with less complexity), and its model architecture is compatible with the generation of faster results for dynamic object detection. Its efficiency improved by increasing the number of layers (with the complexity of the architecture). Individual views (learning) of YOLOv8n or YOLOv8l achieved the specific benefits for the object detections are faster results (with nanoarchitecture design of YOLOv8n) and improved accuracy levels (with the complex architecture of YOLOv8l). This paper uses the fusion-based multiview deep learning techniques, in which the weights of YOLOv8n (nano) and YOLOv8l(large) are fused for proposed model training. The proposed fused model pseudocode is shown in the following algorithm.

Algorithm: Fusion-based Multiview Deep Learning

 # Load the trained model weights
weights_n = torch.load("path_to_yolov8n_weights.pt")
weights_l = torch.load("path_to_yolov8l_weights.pt")

```
# Ensure both models have the same keys
assert weights_n.keys() == weights_l.keys()

# Create a new model for fusion
model_fusion = YOLO("yolov8n.yaml") # Use a base model to start with

# Average the weights
fused_weights = {}
for key in weights_n.keys():
fused_weights[key] = (weights_n[key] + weights_l[key]) / 2

# Load the fused weights into the new model
model_fusion.model.load_state_dict(fused_weights)

# Save the fused model
torch.save(model_fusion.model.state_dict(), "path_to_fused_model_
weights.pt")

# Train the fused model and get its history
history_fusion = train_model(model_fusion, "config.yaml", 1)
```

The proposed Multiview deep learning is implemented with the two views of finetuned weights of the YOLOv8n and YOLOv8l models. It is necessary to maintain the same keys to access the weights of these YOLO models. With the same keys, it is easy to access the final weights of the models. These multiple weights of the models are finally perceived by taking the average weighted values in the fused implementations of the proposed Multiview deep learning technique. The base model for this proposed fusion is YOLOv8n (nano) – it is initiated and derives the model parameters in the object data's first view (or initial learning). This learning stage performs the training with the nano architecture of the YOLOv8n model with fewer levels and less complex architecture. In the fused implementation, average weights are computed by accessing the weights of YOLOv8n (nano) and YOLOv8l (large) to refine the training and its accurate validation results. This fusion is the average weighted fusion YOLO model. The features of object data are extracted, and learning is effectively done by the fused weighted of Multi YOLO v8-nano and YOLOv8-large models. The first view (or learning) of the YOLOv8n model is fused with another view (or learning) YOLOv8l model by passing the fused averaged weighted parameters. In this fused model, the best training is performed without increasing the number of layers or the complexity of YOLOv8n architecture. Fused features of objects are extracted and trained with a fusion of weights of different YOLO-nano

and YOLO-large architectures. The proposed YOLOv8Fusion combines the weights of YOLOv8n (nano) and YOLOv8l (large), combining the strengths of both. YOLOv8n offers fast detection using a lightweight architecture, whereas YOLOv8l offers higher accuracy through its complex design. Thus, by averaging the trained weights of both models, YOLOv8Fusion finds an appropriate balance between speed and accuracy to optimize performance for real-time applications. The complexity of architecture remains less in fusion-based multiview deep models than in YOLOv8-large. Thus, it generates faster results of object detection in video surveillance applications. The weights of YOLOv-large are considered in the fused model. Thus, it generates faster objection results with more accuracy than a single view (or learning) of either YOLOv8-nano or YOLOv8-large models. Experiments were conducted on real-time video data to recognize crime objects with a proposed fusion deep learning model. Its results and comparative study are described in the following section.

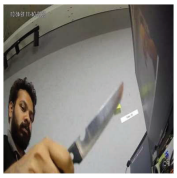
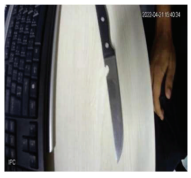
4 Results Analysis and Comparative Study

The suspected objects data is publicly available in [35]-includes faces, knives, and pistol data.

These multi-classified objects are used in the experimental work for the deep result analysis of existing YOLOv8 models and the comparative study of existing YOLO and proposed fusion models. Sample frames of the dataset are shown in Table 2. The training and testing frames are collected from 'Faces-Knifes-Pistols.7vi.yolov8.zip' with sizes of 10319 and 881, respectively. These images are resized according to the YOLOv8 models and fed into other convolution layers; the output is derived from the fully connected layers. Massive training data frames learned using the variants of YOLOv8n, YOLOv8s, and YOLOv8l. Testing data frames are used to evaluate existing YOLO models and proposed fusion-based multiview deep learning models. Results of two distinct fusion-based Multiview models are analyzed with the different performance measures, box loss, classification loss (cls), distribution focus loss (dfl-loss) [36], precision, and recall [37].

Model performance computed with an IoU threshold value of 0.5 is referred to as mAP50, and it is computed with an IoU threshold value from 0.5 to 0.95 by incremented step value of 0.05, which is referred to as mAP50-95. Both mAP50 and mAP50-95 [35] are also vital performance measures for evaluating YOLO models for object detection. YOLOv8 with nano, small, and large are trained with 10319 frames using 26 epochs using the GPU in

Table 2 Sample frames of faces, knifes, and pistols dataset used for the experimental analysis

Sample Frames of Single Objects	Two Objects Detection Sample Frames	Three Objects Detection Sample Frames
		
		
		

Google colab environment. The testing of YOLOv8 models, three existing models with nano, small, and large, and two proposed fusion models, weights level fusion and features level fusion, are experimented with to evaluate the performance. The box value is greatly reduced compared to classification and distribution focus loss at each new epoch. In all these cases, the loss value is gradually reduced when reaching the last 26th epoch in the YOLOv8 models validation and training datasets. The same observation is made experimentally for the YOLOv8n (nano) model, and it is shown in Figure 3. For the ten epochs of another batch of training datasets, loss values, and other performance parameters of YOLOv8s are shown in Figure 4. The precision and recall values are improved for every iterated epoch in training. The mAP50 and mAP50-95 with bounding box overlapping values are optimized at the last epochs in the training.

Multiview weighted deep YOLOv8Fusion retrieves the finetuned weights from both YOLOv8n and YOLOv8l. The YOLOv8Fusion architecture computes the final weight values of YOLOv8-large architecture. Fuse the weights of multiviews of YOLOv8n and YOLOv8l in the proposed YOLOV8Fusion. Fused weight values and efficient training of objects are done to achieve

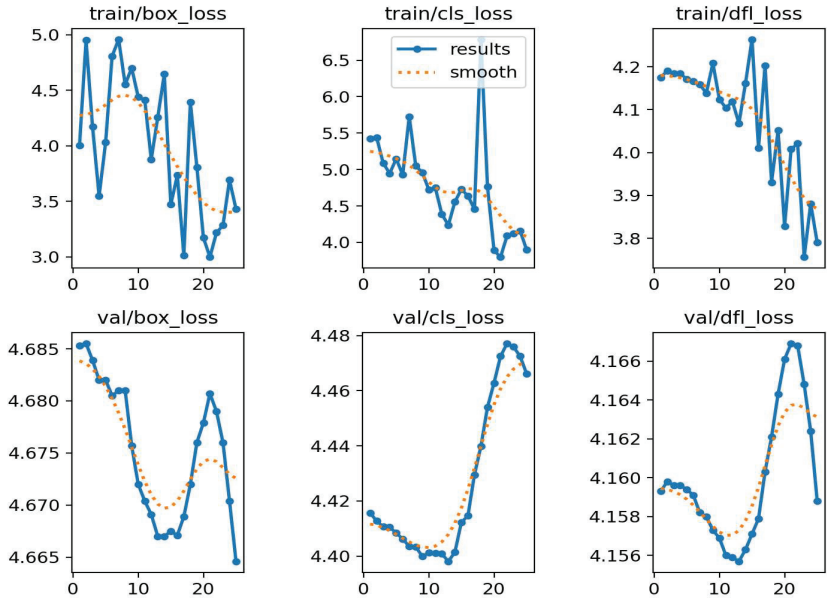


Figure 3 Optimizes the loss value for generations of epoch for training data and validation data.

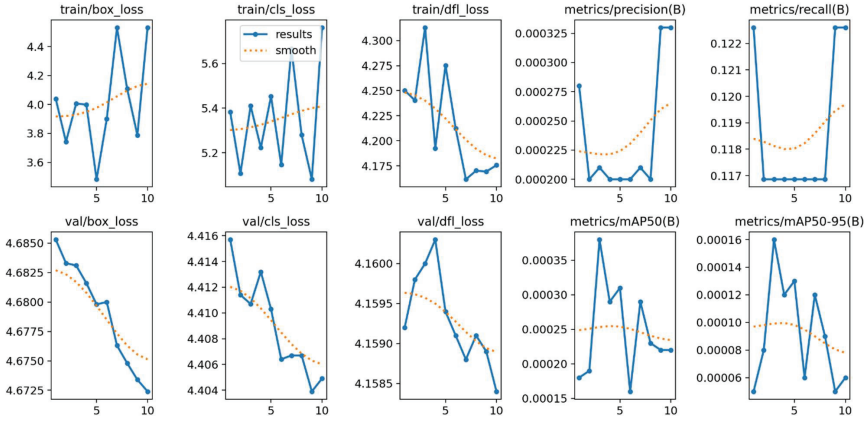


Figure 4 Optimizes the loss and other evaluation parameters for YOLOv8s model.

accurate object detection. Precision and recall values are derived from its confusion matrix results. Figures 5 and 6 depict the comparative analysis of obtained precision and recall values in the evaluation of YOLOv8-nano, YOLOv8-small, YOLOv8-large (three existing models), and proposed

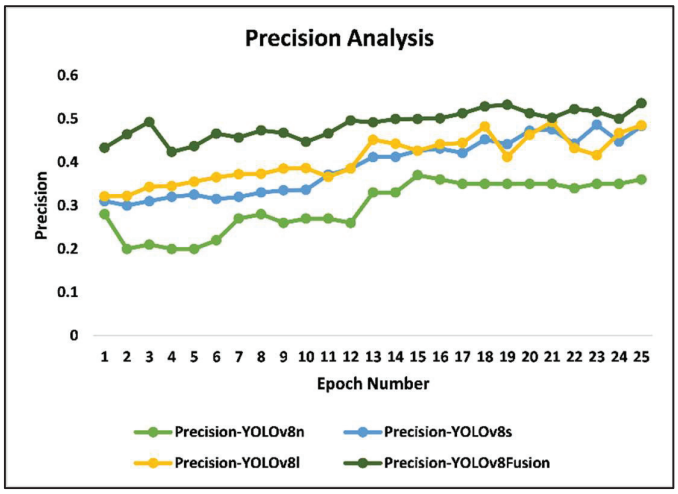


Figure 5 Precision comparative analysis.

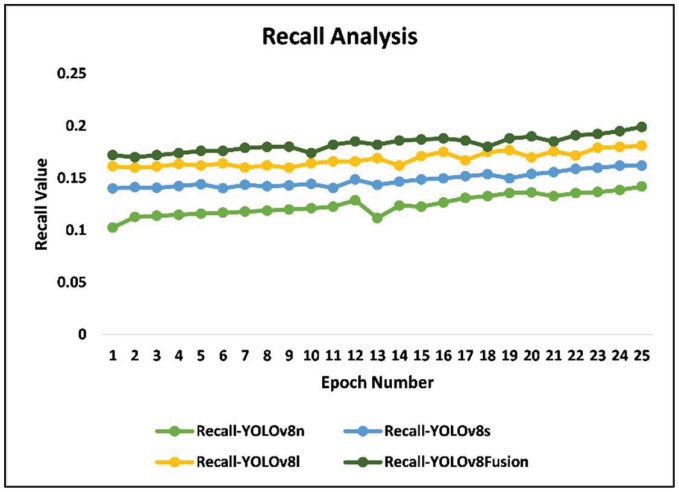


Figure 6 Recall comparative analysis..

YOLOv8Fusion. Obtained precision and recall values are high in the proposed YOLOv8Fusion compared to other YOLOv8 models to the object's detection for the validation datasets. Figure 7 depicts the loss values, and it is noted that the proposed model optimizes the loss more compared to other YOLO models. Further, the proposed work limits the size of video datasets for the training concerning the scalability parameter. Testing the YOLOv8

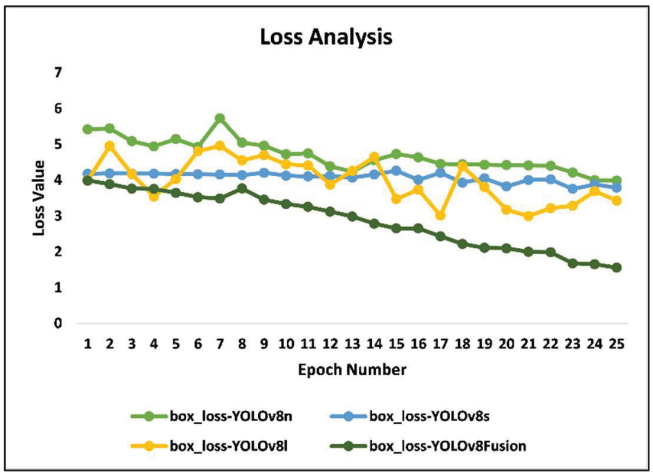


Figure 7 Loss comparative analysis.

models (nano, small, large), and the proposed fusion models for weight-level and feature-level fusion, revealed a consistently declining loss values trend until 26 epochs. Although the box loss significantly indicated better bounding box accuracy against classification and distribution focus losses, the models presented performed best in feature-level fusion as it utilizes complementary features for achieving higher accuracy rates and further optimizing the values of loss metrics. The YOLOv8-nano architecture is faster but has moderate accuracy. The proposed YOLOv8Fusion model improves this by combining the speed of YOLOv8-nano with the precision of YOLOv8-large using fused weight values to achieve optimal speed and accuracy for improved object detection. The challenging problem in future work is addressing the scalability complexity of dynamic object detection with massive video data. Dynamic and effective sampling algorithms must be developed to address object detection problems over massive video data training with YOLO. With regard to huge volumes of video data, it challenges dynamic object detection in the context of scalability. With these kinds of large video data, processing efficiency and high detection accuracy across different scenarios cannot be preserved by current YOLO models. To address such issues, dynamic and efficient sampling algorithms need to be formulated. These algorithms will therefore optimize the selection of samples toward training with reduced redundancy in feature learning. Algorithms able to tackle scalability will permit the processing of large sets within YOLO models for better performance in complex diverse applications through robust video inputs.

5 Conclusion

Detecting suspicious objects by live CCTV is an immense way of finding crimes instantly and saving the public. The state-of-the-art YOLO models accurately perform object classification or detections for high-quality video surveillance systems. The latest YOLOv8 models are designed with three variants: nano, small, and large, based on the different complexity levels of YOLO Darknet architectures. YOLOv8l is robust, and it generates extremely high accurate precision and recall object detection results. YOLOv8-nano is faster; however, its moderate complexity architecture recognizes the object's detection with normal accuracy levels. This paper develops the fusion-based YOLO model, and its design is similar to YOLOv8-nano. It takes fused weight values of Multiview nano and large YOLO models, passed to YOLOv8-nano in the proposed YOLOv8Fusion. The fusion-based Multiview deep learning technique (YOLOv8Fusion) improves the precision with a rate of 4% to 6% and recall with a rate of 2% to 3.5%, and loss is significantly reduced compared with other YOLOv8 models.

Acknowledgments

The authors extend their sincere appreciation to the Centre for Research Projects (CRP), CHRIST (Deemed to be University), Bangalore Central Campus, Bangalore, India. Their generous provision of Seed Money and Grants as Incentive for High Scopus Publication for the academic year 2023-24 (Project Number CU: CRP: SMSS-2340) greatly supported this research endeavor.

References

[1] S. L and C. S. Christopher, "Video Surveillance using Deep Learning – A Review," *2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC)*, Nagercoil, India, 2019, pp. 1-5, doi: 10.1109/ICRAECC43874.2019.8995084.

[2] Upreti Kamal, Peng Sheng-Lung, Kshirsagar Pravin Ramdas, Chakrabarti Prasun, Al-Alshaikh Halah A., Sharma, A. K., Poonia Ramesh Chandra, (2023) A multi-model unified disease diagnosis framework for cyber healthcare using IoMT- cloud computing networks, Journal of

Discrete Mathematical Sciences and Cryptography, 26:6, 1819–1834, DOI: 10.47974/JDMSC-1831.

[3] Y. Lin, Z. Ning, J. Liu, M. Zhang, P. Chen and X. Yang, "Video steganography network based on 3DCNN," *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, Chengdu, China, 2021, pp. 178–181, doi: 10.1109/DSInS54396.2021.9670614.

[4] T. Akilan, Q. J. Wu, A. Safaei, J. Huo and Y. Yang, "A 3D CNN-LSTM-Based Image-to-Image Foreground Segmentation," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 959–971, March 2020, doi: 10.1109/TITS.2019.2900426.

[5] Aggarwal, D., Mittal, S., Upreti, K., and Nayak, P. (2024). Reward Based Garbage Monitoring and Collection System Using Sensors. Journal of Mobile Multimedia, 20(02), 391–410. https://doi.org/10.13052/jmm1550-4646.2026.

[6] Diwan, T., Anirudh, G. and Tembhurne, J.V. Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimed Tools Appl* 82, 9243–9275 (2023). https://doi.org/10.1007/s11042-022-13644-y.

[7] Upreti, K., Singh, P., Jain, D. et al. Progressive loss-aware fine-tuning stepwise learning with GAN augmentation for rice plant disease detection. Multimedia Tools Appl (2024). https://doi.org/10.1007/s11042-024-19255-z.

[8] Hermens, F. Automatic object detection for behavioural research using YOLOv8. *Behav Res* (2024). https://doi.org/10.3758/s13428-024-02420-5.

[9] Elhanashi, A., Dini, P., Saponara, S. et al. TeleStroke: real-time stroke detection with federated learning and YOLOv8 on edge devices. *J Real-Time Image Proc* 21, 121 (2024). https://doi.org/10.1007/s11554-024-01500-1.

[10] P. Rathore, D. Kumar, S. Rajasegarar, M. Palaniswami and J. C. Bezdek, "A Scalable Framework for Trajectory Prediction," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3860–3874, Oct. 2019, doi: 10.1109/TITS.2019.2899179.

[11] Gündüz, M.Ş., Işık, G. A new YOLO-based method for real-time crowd detection from video and performance analysis of YOLO models. *J Real-Time Image Proc* 20, 5 (2023). https://doi.org/10.1007/s11554-023-01276-w.

[12] Yang, L., Chen, G. and Ci, W. Multiclass objects detection algorithm using DarkNet-53 and DenseNet for intelligent vehicles. *EURASIP J.*

Adv. Signal Process. 2023, 85 (2023). https://doi.org/10.1186/s13634-023-01045-8.

[13] Upreti, K., Kapoor, A., Hundekari, S., Upreti, S., Kaul, K., Kapoor, S., and Tiwari, A. (2024). Deep Dive Into Diabetic Retinopathy Identification: A Deep Learning Approach with Blood Vessel Segmentation and Lesion Detection. *Journal of Mobile Multimedia*, 20(02), 495–524. https://doi.org/10.13052/jmm1550-4646.20210.

[14] X. Ni, Z. Ma, J. Liu, B. Shi and H. Liu, "Attention Network for Rail Surface Defect Detection via Consistency of Intersection-over-Union(IoU)-Guided Center-Point Estimation," in *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1694–1705, March 2022, doi: 10.1109/TII.2021.3085848.

[15] Rajendra Prasad, K., Mohammed, M. and Noorullah, R.M. Visual topic models for healthcare data clustering. *Evol. Intel.* 14, 545–562 (2021). https://doi.org/10.1007/s12065-019-00300-y.

[16] Umamakeswari, A., Angelus, J., Kannan, M., Rashikha, Bragadeesh, S.A. (2020). Action Recognition Using 3D CNN and LSTM for Video Analytics. In: Bhateja, V., Satapathy, S., Zhang, YD., Aradhya, V. (eds) Intelligent Computing and Communication. ICICC 2019. Advances in Intelligent Systems and Computing, vol 1034. Springer, Singapore. https://doi.org/10.1007/978-981-15-1084-7_51.

[17] Duarte, F.F., Lau, N., Pereira, A., Reis, L.P. (2024). Study on LSTM and ConvLSTM Memory-Based Deep Reinforcement Learning. In: Rocha, A.P., Steels, L., van den Herik, J. (eds) Agents and Artificial Intelligence. ICAART 2023. Lecture Notes in Computer Science, vol. 14546. Springer, Cham. https://doi.org/10.1007/978-3-031-55326-4_11.

[18] Bhatt, C., Kumar, I., Vijayakumar, V. et al. The state of the art of deep learning models in medical science and their challenges. *Multimedia Systems* 27, 599–613 (2021). https://doi.org/10.1007/s00530-020-00694-1

[19] Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik. 29(2), 102–127 (2019)

[20] Dev, Krishna, Zubair Ashraf, Pranab K. Muhuri, and Sandeep Kumar. Deep autoencoder based domain adaptation for transfer learning. *Multimedia Tools and Applications* 81, no. 16, 22379–22405 (2022).

[21] Sirisha, U., Praveen, S.P., Srinivasu, P.N. et al. Statistical Analysis of Design Aspects of Various YOLO-Based Deep Learning Models for

Object Detection. *Int J Comput Intell Syst* 16, 126 (2023). https://doi.org/10.1007/s44196-023-00302-w

[22] Ningthoujam, R., Pritamdas, K., Singh, L.S. (2024). Comparative Study on YOLOv2 Object Detection Based on Various Pretrained Networks. In: Swain, B.P., Dixit, U.S. (eds) Recent Advances in Electrical and Electronic Engineering. ICSTE 2023. Lecture Notes in Electrical Engineering, vol 1071. Springer, Singapore. https://doi.org/10.1007/978-981-99-4713-3_18.

[23] Zhao, B., Xie, N., Ge, J., Chen, W. (2023). Development of Object Identification APP Based on YoloV2. In: Atiquzzaman, M., Yen, N., Xu, Z. (eds) Proceedings of the 4th International Conference on Big Data Analytics for Cyber-Physical System in Smart City – Volume 1. BDCPS 2022. Lecture Notes on Data Engineering and Communications Technologies, vol. 167. Springer, Singapore. https://doi.org/10.1007/978-981-99-0880-6_5.

[24] Tyagi, B., Nigam, S., Singh, R. (2023). Person Detection Using YOLOv3. In: Kumar, R., Verma, A.K., Sharma, T.K., Verma, O.P., Sharma, S. (eds) Soft Computing: Theories and Applications. Lecture Notes in Networks and Systems, vol 627. Springer, Singapore. https://doi.org/10.1007/978-981-19-9858-4_77.

[25] Guo, B., Wang, H., Jin, L. et al. DCM3-YOLOv4: A Real-Time Multi-Object Detection Framework. *Automot. Innov.* 7, 283–299 (2024). https://doi.org/10.1007/s42154-023-00258-9.

[26] Arkin, E., Yadikar, N., Xu, X. et al. A survey: object detection methods from CNN to transformer. *Multimed Tools Appl* 82, 21353–21383 (2023). https://doi.org/10.1007/s11042-022-13801-3.

[27] Cao, W., Li, T., Liu, Q. et al. PANet: Pluralistic Attention Network for Few-Shot Image Classification. *Neural Process Lett* 56, 209 (2024). https://doi.org/10.1007/s11063-024-11638-5.

[28] Nakhodnov, M.S., Kodryan, M.S., Lobacheva, E.M. et al. Loss Function Dynamics and Landscape for Deep Neural Networks Trained with Quadratic Loss. *Dokl. Math.* 106 (Suppl 1), S43–S62 (2022).

[29] Parulekar, B., Singh, N. and Ramiya, A.M. Evaluation of segment anything model (SAM) for automated labelling in machine learning classification of UAV geospatial data. *Earth Sci Inform* (2024). https://doi.org/10.1007/s12145-024-01402-7.

[30] Boehme, M.G., Al-Turjman, F. (2024). Enhancing Object Detection Capabilities: A Comprehensive Exploration and Finetuning of YOLOv5 Algorithm Across Diverse Datasets. In: Al-Turjman, F. (eds) The

Smart IoT Blueprint: Engineering a Connected Future. AIoTSS 2024. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-031-63103-0_9.

[31] Gupta, C., Gill, N.S., Gulia, P. et al. A novel finetuned YOLOv6 transfer learning model for real-time object detection. *J Real-Time Image Proc* 20, 42 (2023). https://doi.org/10.1007/s11554-023-01299-3.

[32] Luo, X. et al. (2024). Improved YOLOv7-Tiny Insulator Defect Detection Based on Drone Images. In: Huang, DS., Zhang, X., Guo, J. (eds) Advanced Intelligent Computing Technology and Applications. ICIC 2024. Lecture Notes in Computer Science, vol. 14866. Springer, Singapore. https://doi.org/10.1007/978-981-97-5594-3_29.

[33] Zhao, H., Zhou, Y., Zhang, L., Peng, Y., Hu, X., Peng, H. and Cai, X.. Mixed YOLOv3-LITE: A lightweight real-time object detection method. *Sensors*, 20(7), p. 1861 (2020).

[34] Edmundo Casas, Leo Ramos, Cristian Romero, Francklin Rivas-Echeverría, A comparative study of YOLOv5 and YOLOv8 for corrosion segmentation tasks in metal surfaces, Array, Volume 22, 2024, 100351, ISSN 2590-0056.

[35] https://universe.roboflow.com/cigarettesmokingdetection/faces-knifes-pistols.

[36] Xiao, B., Nguyen, M. and Yan, W.Q. Fruit ripeness identification using YOLOv8 model. *Multimed Tools Appl* 83, 28039–28056 (2024). https://doi.org/10.1007/s11042-023-16570-9.

[37] Simeth, A., Kumar, A.A. and Plapper, P. Flexible and robust detection for assembly automation with YOLOv5: a case study on HMLV manufacturing line. *J Intell Manuf* (2024). https://doi.org/10.1007/s10845-024-02411-5.

Biographies

Ramesh Chandra Poonia is a Professor in the Department of Computer Science at CHRIST (Deemed to be University), NCR Delhi Campus, India. He is internationally recognized for his research in sustainable technologies, focusing on energy-efficient algorithms, cyber-physical systems, and computational intelligence, particularly machine learning and data analytics. His contributions rank him among the top 2% of scientists worldwide, as recognized by Stanford University and Elsevier. Dr. Poonia has completed two distinguished postdoctoral fellowships: one at the Cyber-Physical Systems Laboratory at the Norwegian University of Science and Technology (NTNU) in Norway, and another through an international collaboration aimed at predicting pandemic diseases using machine learning, involving Oakland University in the USA and Imam University in Saudi Arabia. He earned his Ph.D. in Computer Science from Banasthali University, India, in 2013, and an M.Tech. in Data Science and Engineering from the Birla Institute of Technology and Science (BITS), Pilani, India. With an extensive portfolio of research, Dr. Poonia has served as lead editor for numerous special issues, books, and proceedings with renowned publishers such as Springer, Taylor & Francis, and Elsevier. He is also an associate editor for the Journal of Sustainable Computing: Informatics and Systems (Elsevier) and serves as a series editor for Computational and Intelligent Systems (CRC Press). As the founder of the SUSCOM and ICSCPS conferences, he received the 2024 Research Innovation Award for his outstanding contributions to the advancement of sustainable and intelligent systems.

Kamal Upreti is currently working as an Associate Professor in Department of Computer Science, CHRIST (Deemed to be University), Delhi NCR, Ghaziabad, India. He completed is B. Tech (Hons) Degree from UPTU, M. Tech (Gold Medalist), PGDM (Executive) from IMT Ghaziabad and PhD in Department of Computer Science & Engineering. He has completed Postdoc from National Taipei University of Business, TAIWAN funded by MHRD.

He has published 50+ Patents, 32+Magazine issues and 113+ Research papers in in various reputed Journals and international Conferences. His areas of Interest such as Modern Physics, Data Analytics, Cyber Security, Machine Learning, Health Care, Embedded System and Cloud Computing. He has published more than 45+ authored and edited books under CRC Press, IGI Global, Oxford Press and Arihant Publication. He is having enriched years' experience in corporate and teaching experience in Engineering Colleges.

He worked with HCL, NECHCL, Hindustan Times, Dehradun Institute of Technology and Delhi Institute of Advanced Studies, with more than 15+ years of enrich experience in research, Academics and Corporate. He also worked in NECHCL in Japan having project – "Hydrastore" funded by joint collaboration between HCL and NECHCL Company. Dr. Upreti worked on Government project – "Integrated Power Development Scheme (IPDS)" was launched by Ministry of Power, Government of India with the objectives of Strengthening of sub-transmission and distribution network in the urban areas. He has completed work with Joint collaboration with GB PANT & AIIMS Delhi, under funded project of ICMR Scheme on Cardiovascular diseases prediction strokes using Machine Learning Techniques from year 2017–2020 of having fund of 80 Lakhs. He got 5 Lakhs fund from DST SERB for conducting International Conference, ICSCPS-2024, 13–14 Sept 2024. Recently, he got 10 Lakhs fund from AICTE – Inter-Institutional Biomedical Innovations and Entrepreneurship Program (AICTE-IBIP) for

2024–2026. He has attended as a Session Chair Person in National, International conference and key note speaker in various platforms such as Skill based training, Corporate Trainer, Guest faculty and faculty development Programme. He awarded as best teacher, best researcher, extra academic performer and Gold Medalist in M. Tech programme.

Nidhi Singh is an accomplished academician with over 14 years of experience spanning teaching, training, and corporate sectors. Currently serving as an Assistant Professor at G.D. Goenka University in Haryana, India, she holds a Ph.D. in Management and an MBA in Information Technology and Marketing. Her expertise covers business analytics, information technology, and emerging fields such as machine learning, data visualization, and disruptive technologies in higher education. She has published over twenty research papers in reputed journals like SCOPUS, WOS, and ABDC. Dr. Singh also holds certifications from prestigious institutions, including Harvard Business School and the Indian Institute of Management Visakhapatnam, further solidifying her expertise in emerging technologies and business strategies.

Jyoti Kesarwani is an accomplished academic professional with a strong background in Computer Applications. I hold a Master of Computer

Applications (MCA) and pursuing Ph.D. in the domain of computer science. Currently, I serve as an Assistant Professor at United College of Engineering and Research. I taught many subjects such as C Programming, Design and Analysis of Algorithms (DAA) and Artificial Intelligence (AI). I have developed a keen interest in the fields of Machine Learning (ML) and Deep Learning (DL). My dedication to education and research continues to inspire and shape the next generation of computer science profession.

Mohammad Shabbir Alam is presently working as Senior Lecturer in College of Engineering and Computer Science, Jazan University (Public University), Jazan, Kingdom of Saudi Arabia. He received his Master in Computer Science & Applications (MCA) in years 2007 from Aligarh Muslim University, Aligarh, India. More than 15 years of academic and industry experiences in area of Computer Science and Information to Technology.

He has published 1 UK Patents, 2 German Patents and 4 Australian patents, 4 Books, 1 Book chapter and more than 30+ research papers in reputed international journals and national/international conference proceedings. He is an author of Data structure and Algorithm book. His areas of research interest include Deep learning, Blockchain, Machine Learning and Health Care.