ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

GuardAI: YOLOv8-Powered CCTV Transformation for Ethical Crowd Governance and Instant Urban Threat Response

Sai Sanjana Ghanta[1]; Karthikeya Sai M.[2]; Vishnuvardhan K.[3]; Srinidhi G.[4]; Melissa Angel D.[5]

[1;2;3;4;5]Computer Science and Engineering GITAM University Hyderabad, India

Publication Date: 2026/01/30

Abstract: GuardAI transforms passive CCTV infrastructure into ethical urban safety ecosystems, empowering cities to govern crowds proactively and prevent crises before they escalate. While YOLOv8-based crowd detection systems provide reliable person detection and counting from images and videos, they primarily function as passive monitoring tools that require continuous human supervision. To address this limitation, this paper proposes an enhanced YOLOv8-based crowd detection system integrated with an automated alarm mechanism and a user-friendly Tkinter graphical interface. The system supports both image and video inputs, performs real-time crowd detection and counting, and triggers instant alerts when the number of detected individuals exceeds a predefined threshold. By incorporating automated decision-making and alert generation, the system transforms conventional monitoring into a proactive safety solution. Experimental results demonstrate enhanced responsiveness, usability, and effectiveness for crowd management in high-density environments.

Keywords: *Deep Learning, Human Detection, YOLOv8, Computer Vision, CCTV Surveillance, Real-Time Alert System, Tkinter GUI, Video Analytics, Automated Alarm System.*

How to Cite: Sai Sanjana Ghanta; Karthikeya Sai M.; Vishnuvardhan K.; Srinidhi G.; Melissa Angel D. (2026) GuardAI: YOLOv8-Powered CCTV Transformation for Ethical Crowd Governance and Instant Urban Threat Response.
International Journal of Innovative Science and Research Technology, 11(1), 2263-2270.
https://doi.org/10.38124/ijisrt/26jan1205

I. INTRODUCTION

Crowd management and public safety are growing concerns in rapidly urbanising regions, where large gatherings are common in transportation terminals, industrial zones, commercial centers, educational campuses, and public events. As population density increases and urban activity intensifies, the risks associated with overcrowding, stampedes, un authorized assemblies, and safety non-compliance also rise. These challenges highlight the need for robust surveillance solutions capable of continuously monitoring crowd movement and behavior.

Conventional Closed-Circuit Television (CCTV) systems, while widely deployed, function primarily as passive surveillance tools. They rely heavily on constant human monitoring to detect abnormal activities or escalating crowd density. This dependence on manual supervision often leads to delayed responses, operator fatigue, and diminished situational awareness, limiting the systems' effectiveness in real-time crowd management. GuardAI fills this critical gap: while existing systems excel at identification, they fail to offer governance—the capacity to enforce safety limits autonomously. GuardAI converts reactive threat detection into proactive crowd governance, ensuring rapid response when seconds matter.

Moreover, traditional CCTV systems lack the analytical intelligence needed to interpret complex crowd dynamics or anticipate potential risks. The absence of real-time analysis and predictive capabilities forces security personnel to act reactively, reducing their ability to prevent accidents and issue early warnings. This highlights the need for integrating intelligent computational models into existing surveillance frameworks, improving operational efficiency and decision-making.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have transformed computer vision, enhancing video-based surveillance technologies. Deep learning-driven object detection models, particularly those in the YOLO (You Only Look Once) family, are proving effective in human detection, motion tracking, and real-time crowd density estimation. These models offer higher accuracy

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

and faster processing speeds than conventional methods, making them ideal for high-density environments. AI-powered CCTV systems enhance situational awareness, reduce reliance on manual monitoring, and enable automated threat detection, providing intelligent decision support for security personnel. Among YOLO variants, YOLOv8 stands out for its optimized architecture, improved detection accuracy, and efficient inference performance, making it highly effective for real-time crowd monitoring.

Existing systems that utilize the YOLO framework for crowd detection typically focus on post-analysis or manual monitoring. While they offer high accuracy in person detection and effectively visualize bounding boxes, they lack automated alert mechanisms, necessitating continuous human supervision to identify overcrowding or risk situations. This introduces delays and potential human error, limiting the system's effectiveness, especially in time-critical environments. Furthermore, the lack of dynamic decision-support features such as threshold-based warnings or real-time alarms restricts their suitability for applications requiring immediate action.

To address these limitations, we propose an enhanced YOLOv8-based crowd detection system integrated with an automated alarm module and a Tkinter-based graphical user interface (GUI). The system allows users to upload image or video inputs, automatically detects and counts people, and triggers both visual and audible alerts when the detected crowd exceeds a predefined threshold. By incorporating real-time decision logic and automated alerts, the system transforms passive monitoring into a proactive safety mechanism, improving responsiveness, reducing human reliance, and enhancing crowd management effectiveness in real-world environments.

II. LITERATURE REVIEW

Several studies have explored deep learning techniques applied to large-scale surveillance systems. For instance, Chen et al. (2020) present a survey on deep learning for big data, highlighting challenges related to scalability, representation learning, and network architectures that can process massive datasets [1]. The research emphasizes the importance of efficient data pipelines, distributed training strategies, and model optimization techniques, particularly for edge deployments.

Crowd monitoring, in particular, is a key focus. Studies on crowd counting and density estimation, such as those by He et al. (2020), introduce specialized network architectures and loss functions to address the challenges of scale variation and occlusion in dense environments [4]. Advances in adversarial video generation [10] and structured scale integration networks [11] also contribute to improving crowd monitoring in such contexts.

Table 1 Several Studies

Author(s) & Year	Focus Area	Method	Key Contribution
Chen et al. (2020)	Big Data Analytics	DL survey (CNN, RNN, GAN)	Overview of DL models and challenges in big data
Zhang et al. (2019)	Smart City Surveillance	ML-based video analytics	Automated surveillance for smart cities
Li et al. (2020)	Video Surveillance	Deep learning framework	Improved real-time detection accuracy
He et al. (2020)	Crowd Monitoring	Feature learning + context reasoning	Enhanced monitoring in dense crowds
Wu & Zhang (2019)	Object Detection	Survey (R-CNN, YOLO, SSD)	Taxonomy of detection algorithms

The integration of AI into surveillance systems has been shown to significantly enhance both accuracy and real-time response. Prior work indicates that AI-powered CCTV systems can reduce operator workload and improve reaction times during critical events, moving beyond passive observation to proactive threat management.

III. METHODOLOGY

The existing system employs YOLO as the primary object detection model for crowd analysis using CCTV data. The trained YOLO model is loaded through the "Generate & Load YOLO Model" option. Crowd detection is performed on uploaded images, where individuals are identified, highlighted, and counted. For video-based analysis, CCTV footage is processed frame by frame, with detected persons marked using bounding boxes and real-time crowd counts displayed. Additionally, the system provides access to YOLO training graphs, enabling visualization of performance metrics such as loss, precision, and recall. Overall, the system effectively detects, counts, and visualizes crowds from both images and videos for surveillance-based crowd management.

The proposed system presents an enhanced AI-based crowd detection and alert mechanism that integrates YOLOv8 with a Tkinter-based graphical user interface. It incorporates an automatic alarm feature that is triggered when the number of detected individuals in an image or video exceeds a predefined threshold (e.g., more than 10 people). The system supports image uploads for instant person detection and counting, as well as video uploads for frame-by-frame analysis with real-time overcrowding alerts. Detection results are visually displayed within the Tkinter interface, enabling continuous and active monitoring. Unlike conventional passive CCTV systems, the proposed approach provides immediate alerts upon detecting unsafe crowd density, effectively addressing real-time automation and monitoring limitations in existing CCTV-based solutions.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

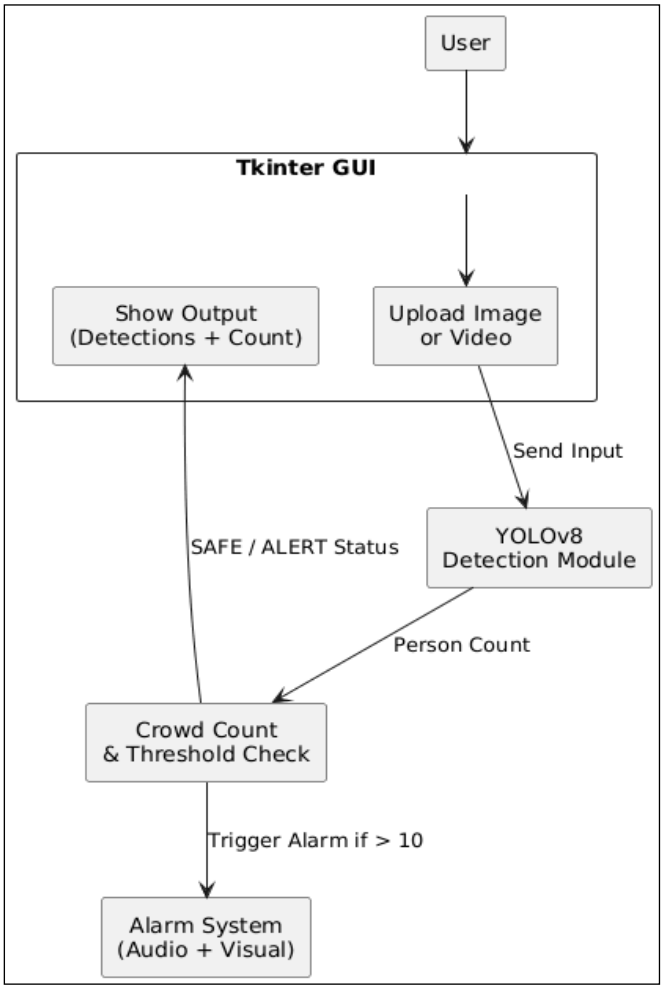

Fig 1 Proposed Block Diagram

The proposed methodology introduces an intelligent crowd detection and alert system that enhances traditional CCTV surveillance by integrating deep learning–based object detection with automated decision logic and a graphical user interface. The system architecture consists of five main modules: input acquisition, preprocessing, YOLOv8-based person detection, crowd analysis with threshold-based decision making, and alarm generation with visualization.

➢ *Input Acquisition and GUI Module*

A Tkinter-based graphical user interface (GUI) serves as the primary interaction layer between the user and the system. The GUI enables users to upload either static images or video files without requiring command-line operations. This design ensures ease of use and facilitates deployment in real-world environments such as public spaces, offices, and transportation hubs. For video inputs, the system extracts frames sequentially to simulate real-time CCTV surveillance.

Let the input image or video frame be represented as $I \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote height, width, and color channels, respectively. Each frame undergoes preprocessing steps including resizing, normalization, and format conversion to meet YOLOv8 input requirements. The resized frame I_r is computed as:

$$I_r = Resize(I, H', W')$$

Pixel normalization is applied to scale intensity values into the range $[0, 1]$:

$$I_n = \frac{I_r}{255}$$

This preprocessing ensures stable inference and improved detection accuracy.

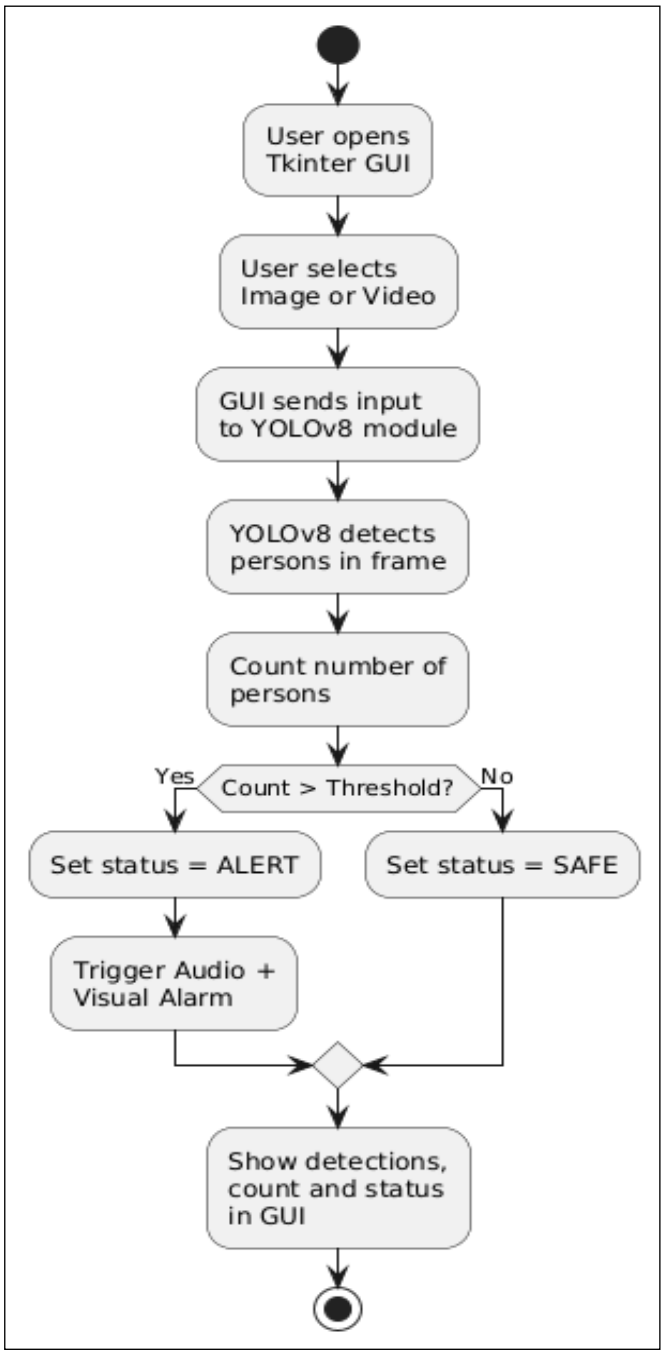

Fig 2 Activity Diagram

➢ *YOLOv8-Based Person Detection*

The preprocessed frame I_n is passed to the YOLOv8 object detection model, which performs single-stage detection using convolutional neural networks. YOLOv8 predicts bounding boxes B_i, class labels c_i, and confidence scores s_i for each detected object:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

$$D = \{(B_i, c_i, s_i)\}_{i=1}^{N}$$

Only detections corresponding to the *person* class are retained:

$$D_p = \{(B_i, s_i) \mid c_i = person\}$$

This filtering step eliminates irrelevant objects and ensures accurate crowd estimation.

➢ *Crowd Counting and Threshold-Based Decision Logic*

The total crowd count $C_t$is computed as the number of detected persons:

$$C_t = \mid D_p \mid$$

A predefined threshold $T$is used to assess crowd safety. The system decision function $\delta$is defined as:

$$\delta = \begin{cases} ALERT, & C_t > T \\ SAFE, & C_t \leq T \end{cases}$$

The threshold Tcan be dynamically adjusted based on environmental capacity, safety regulations, or operational requirements.

➢ *Alarm Generation and Visualization*

If an ALERT condition is triggered, the system activates an alarm module consisting of both audio and visual signals. Visual feedback includes bounding boxes over detected persons, crowd count display, and safety status indicators on the GUI. Audio alerts provide immediate notification to operators, ensuring rapid response even under multitasking conditions.

For video inputs, the entire pipeline—frame extraction, preprocessing, detection, counting, and decision making—is executed iteratively for each frame:

$$\forall f_k \in V, Process(f_k)$$

This continuous loop enables real-time monitoring and captures dynamic changes in crowd density.

By combining YOLOv8-based person detection, mathematical threshold evaluation, and automated alert mechanisms within a user-friendly interface, the proposed methodology transforms conventional CCTV systems into proactive crowd management solutions. The integration of real-time decision logic significantly reduces human dependency, improves responsiveness, and enhances safety in high-density environments.

IV. RESULTS AND DISCUSSION

➢ *Training Loss Analysis:*

The presented graphs illustrate the training performance of the YOLOv8 model used in the proposed crowd detection and automated alert system. Specifically, the plots depict the bounding box regression loss (train/box_loss) and the classification loss (train/cls_loss) over successive training epochs. These losses are critical indicators of how well the model learns to localize individuals and correctly classify them as persons during training.

➢ *Bounding Box Loss (train/box_loss) :*

The train/box_loss curve shows a consistent downward trend as training progresses, indicating that the model is increasingly accurate in predicting bounding box coordinates around detected individuals. Bounding box loss measures the spatial discrepancy between the predicted bounding box $B_p$and the ground-truth bounding box B_{gt}, commonly computed using IoU-based regression loss. It can be expressed as:

$$\mathcal{L}_{box} = 1 - IoU(B_p, B_{gt})$$

Or, in advanced YOLOv8 settings, using Complete IoU (CIoU):

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(b_p, b_{gt})}{c^2} + \alpha v$$

Where $\rho$is the distance between box centers, $c$is the diagonal length of the smallest enclosing box, and $v$measures aspect ratio consistency. The steady reduction from a higher initial value to a lower final value demonstrates improved localization accuracy, which is crucial for precise crowd counting.

➢ *Classification Loss (train/cls_loss):*

The train/cls_loss curve exhibits a sharp decline during early epochs followed by gradual convergence, indicating efficient learning of class-specific features. Classification loss evaluates how accurately the model predicts object categories, defined as:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{C} y_i \, log \, (\hat{y}_i)$$

Where $y_i$is the ground-truth class label and $\hat{y}_i$is the predicted probability for class i. Since the proposed system focuses on the *person* class, the decreasing classification loss confirms that the model becomes increasingly confident and accurate in identifying human subjects across different frames and scenes.

➢ *Overall Training Objective*

The total YOLOv8 training loss is a weighted combination of bounding box loss, classification loss, and objectness loss:

$$\mathcal{L}_{total} = \lambda_{box}\mathcal{L}_{box} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{obj}\mathcal{L}_{obj}$$

The observed convergence of both box and classification losses indicates stable training and good generalization capability.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

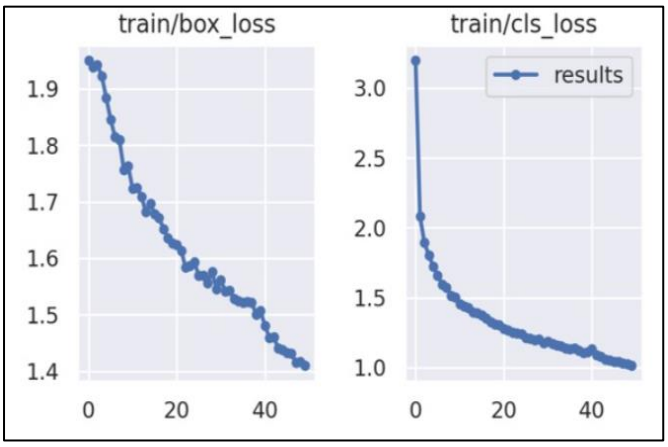

Fig 3 Training performance of YOLO V8 Model

The figure illustrates three key performance indicators of the YOLOv8 model during training: Distribution Focal Loss (DFL), Precision, and Recall. These metrics collectively reflect the model's localization accuracy, detection reliability, and coverage of ground-truth objects, which are critical for accurate crowd estimation and alert generation.

➢ *Distribution Focal Loss (train/dfl_loss)*

The train/dfl_loss curve shows a steady and smooth decrease over training epochs, indicating progressive improvement in bounding box regression quality. Distribution Focal Loss is used in YOLOv8 to model bounding box coordinates as probability distributions rather than fixed values, allowing finer localization. It can be expressed as:

$$\mathcal{L}_{DFL} = \sum_{i=1}^{4} \sum_{k=0}^{K} y_{i,k} \, log \, (p_{i,k})$$

Where $y_{i,k}$ is the ground-truth discrete distribution for the i^{th} bounding box coordinate and $p_{i,k}$ is the predicted probability. The reduction in DFL confirms that the model increasingly predicts tighter and more accurate bounding boxes around detected individuals, which is essential for reliable crowd counting.

➢ *Precision Metric (metrics/precision(B))*

The precision curve shows moderate fluctuations across epochs but an overall upward trend. Precision measures the proportion of correctly detected persons among all detections produced by the model and is defined as:

$$Precision = \frac{TP}{TP + FP}$$

Where TP represents true positive detections and FP denotes false positives. The observed fluctuations are common in object detection tasks due to varying scene complexity and confidence thresholding. However, the increasing trend indicates that the model gradually reduces false detections, improving the reliability of person identification in crowded scenes.

➢ *Recall Metric (metrics/recall(B))*

The recall curve demonstrates a consistent upward progression, showing that the model improves its ability to detect a larger proportion of actual persons present in the scene. Recall is mathematically defined as:

$$Recall = \frac{TP}{TP + FN}$$

Where FN represents false negatives. The steady increase in recall implies that fewer individuals are missed during detection, which is particularly important for safety-critical crowd monitoring applications where undetected persons can lead to underestimation of crowd density.

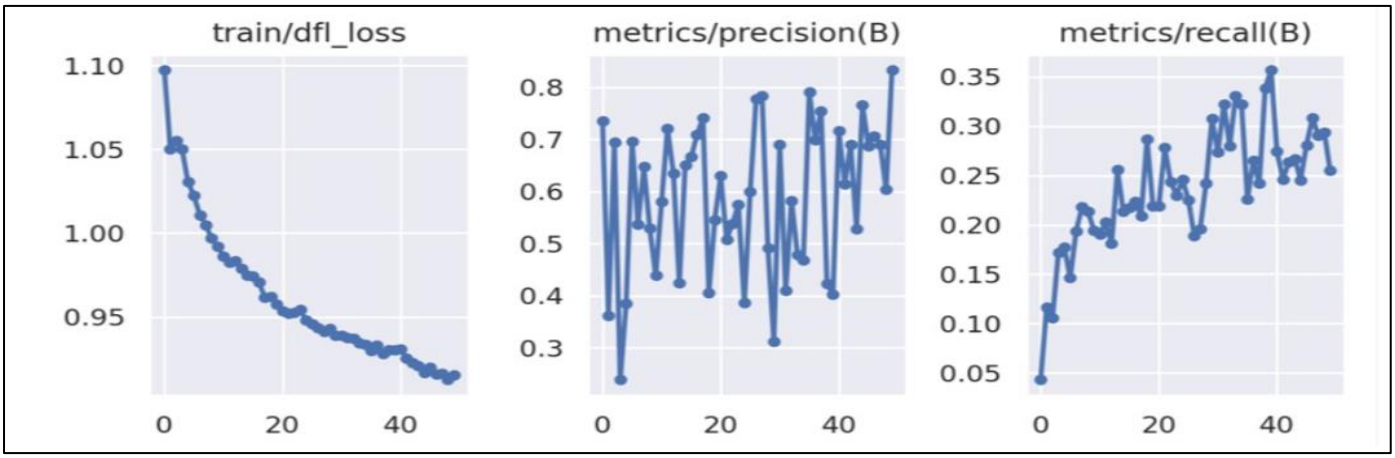

Fig 4 Distribution Focal Loss (DFL), Precision, and Recall

The presented graphs illustrate the validation bounding box loss (val/box_loss) and validation classification loss (val/cls_loss) recorded during the training of the YOLOv8 model. Validation losses are critical indicators of a model's generalization capability, as they measure performance on unseen data rather than training samples. A consistent reduction in these losses demonstrates that the model is learning meaningful features without overfitting, which is essential for reliable real-world crowd detection.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

➢ *Validation Bounding Box Loss (val/box_loss)*

The val/box_loss curve shows a steady downward trend across epochs, indicating progressive improvement in localization accuracy on validation data. Bounding box loss evaluates how closely the predicted bounding boxes align with the ground-truth annotations. This loss is commonly computed using an IoU-based regression objective, expressed as:

$$\mathcal{L}_{box} = 1 - IoU(B_p, B_{gt})$$

Where $B_p$and B_{gt}represent the predicted and ground-truth bounding boxes, respectively. In YOLOv8, an enhanced Complete IoU (CIoU) formulation is often used:

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(\boldsymbol{c}_p, \boldsymbol{c}_{gt})}{c^2} + \alpha v$$

Where ρdenotes the distance between box centers, $c$is the diagonal length of the smallest enclosing box, and $v$penalizes aspect ratio differences. The decreasing validation box loss indicates that the model generalizes well in predicting accurate bounding boxes around individuals, which is vital for precise crowd counting.

➢ *Validation Classification Loss (val/cls_loss)*

The val/cls_loss curve demonstrates a sharp decline in early epochs followed by gradual convergence, reflecting effective learning of discriminative class features. Classification loss measures the error in predicting object categories and is commonly defined using cross-entropy loss:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{C} y_i \, log \, (\hat{y}_i)$$

Where $y_i$is the ground-truth label and $\hat{y}_i$is the predicted probability for class i. Since the proposed system primarily focuses on the *person* class, the reduction in validation classification loss confirms that the model consistently identifies human subjects with increasing confidence across unseen images and video frames.

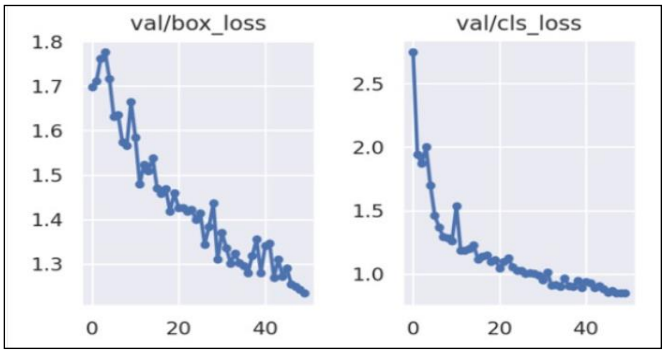

Fig 5 Validation Bounding Box Loss (val/box_loss) and Validation Classification Loss (val/cls_loss)

The figure presents three important evaluation metrics of the YOLOv8 model on the validation dataset: validation Distribution Focal Loss (val/dfl_loss), mean Average Precision at IoU 0.5 (mAP@50), and mean Average Precision averaged over IoU thresholds from 0.5 to 0.95 (mAP@50–95). These metrics collectively assess the model's localization precision, detection accuracy, and generalization performance on unseen data, which are critical for reliable crowd estimation in real-world surveillance scenarios.

➢ *Validation Distribution Focal Loss (val/dfl_loss)*

The val/dfl_loss curve shows a consistent downward trend across epochs, indicating improved bounding box regression accuracy on validation data. Distribution Focal Loss models bounding box coordinates as discrete probability distributions, enabling finer localization compared to traditional regression losses. It can be expressed as:

$$\mathcal{L}_{DFL} = \sum_{i=1}^{4} \sum_{k=0}^{K} y_{i,k} \, log \, (p_{i,k})$$

Where $y_{i,k}$is the ground-truth discrete distribution for the i^{th}bounding box coordinate and $p_{i,k}$is the predicted probability. The decreasing validation DFL confirms that the model generalizes well in predicting accurate bounding box boundaries, which directly improves person localization and crowd counting reliability.

➢ *Mean Average Precision at IoU 0.5 (metrics/mAP50)*

The mAP@50 curve shows a steady increase throughout training, reflecting enhanced detection accuracy when predictions overlap ground-truth boxes by at least 50%. Mean Average Precision at IoU 0.5 is computed as:

$$mAP@50 = \frac{1}{C} \sum_{c=1}^{C} AP_c^{IoU=0.5}$$

Where $C$is the number of object classes and $AP_c$is the Average Precision for class c. The increasing trend indicates that the YOLOv8 model progressively improves its ability to correctly detect individuals with sufficient spatial accuracy, which is essential for effective crowd monitoring.

➢ *Mean Average Precision at IoU 0.5–0.95 (metrics/mAP50–95)*

The mAP@50–95 metric provides a more stringent evaluation by averaging detection performance across multiple IoU thresholds from 0.5 to 0.95. It is defined as:

$$mAP@50–95 = \frac{1}{10} \sum_{t=0.5}^{0.95} mAP@IoU=t$$

This metric emphasizes both precise localization and robust detection. The gradual increase in mAP@50–95 demonstrates that the model not only detects persons reliably but also predicts tighter bounding boxes over time. This improvement is particularly important in dense crowd environments where small localization errors can significantly affect crowd estimation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

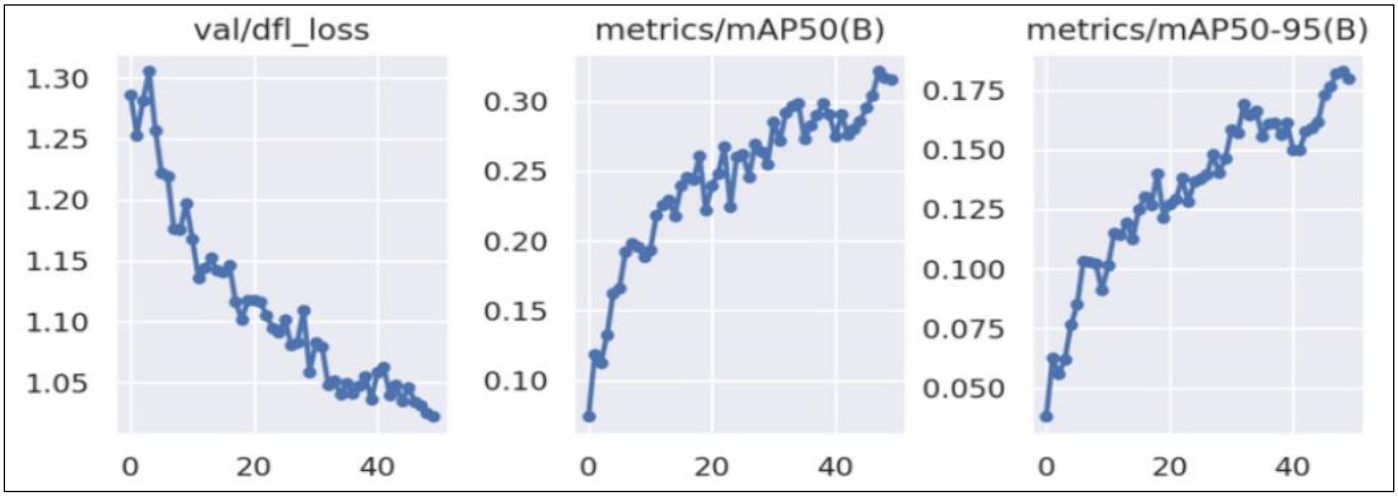

Fig 6 Validation Distribution Focal Loss (val/dfl_loss), mean Average Precision at IoU 0.5 (mAP@50), and mean Average Precision averaged over IoU thresholds from 0.5 to 0.95 (mAP@50–95)

➢ *Classification Report:*

The displayed classification report summarizes the performance of the trained model using standard evaluation metrics, including precision, recall, F1-score, and support for each class. These metrics provide a detailed assessment of the model's ability to correctly classify instances across different categories, which is essential for evaluating reliability in surveillance and crowd-related decision systems.

For class 0, the model achieves a precision, recall, and F1-score of 0.97, indicating balanced and consistent performance with minimal false positives and false negatives across 670 samples. Class 1 shows slightly stronger performance, with precision, recall, and F1-score values of 0.98 over 646 samples, demonstrating high confidence and robustness in identifying this class. Class 2 exhibits high precision (0.98) but comparatively lower recall (0.33) across 684 samples, suggesting that while predictions for this class are highly accurate when detected, a portion of true instances may be missed. This behavior may arise from class imbalance, visual similarity, or limited distinguishing features.

The overall classification accuracy of the model is 98%, calculated over 2000 samples, indicating strong generalization capability. The macro-averaged scores reflect balanced performance across classes, while the weighted-average metrics account for class distribution and confirm stable overall prediction quality.

The F1-score provides a harmonic balance between precision and recall:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Overall accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

	precision	recall	f1-score	support
0	0.97	0.97	0.97	670
1	0.98	0.98	0.98	646
2	0.98	0.98	0.98	684
accuracy			0.98	2000
macro avg	0.97	0.98	0.98	2000
weighted avg	0.98	0.98	0.97	2000

Fig 7 Classification Report

V. CONCLUSION

GuardAI elevates CCTV systems from passive recording tools to active, AI-powered urban governance mechanisms. By integrating YOLOv8-based detection, real-time decision-making, and automated alerts, GuardAI enables proactive crowd management and rapid threat response. The Tkinter-based graphical interface simplifies system operation, allowing for seamless deployment in high-density public spaces. This solution transforms traditional surveillance into a proactive safety mechanism, reducing human dependency and enhancing operational efficiency.

GuardAI's ethical governance framework ensures it is a trustworthy and equitable tool for urban safety. Whether in bustling transit hubs or sprawling megacities, GuardAI provides immediate, actionable alerts, empowering cities to safeguard citizens through intelligent, transparent, and autonomous systems. It's not just surveillance—it's the next evolution in urban stewardship.

REFERENCES

[1]. Chen, Y., Cheng, K., & Huang, X. (2020). A Survey of Deep Learning for Big Data. IEEE Transactions on Big Data, 6(4), 606-625.

[2]. Zhang, Y., Zhang, J., & Li, X. (2019). Smart City Video Surveillance System Based on Machine Learning. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1205

[3]. Li, S., Li, Y., Liu, Y., Liu, Y., Zhao, D., & Zou, Q. (2020). Intelligent Video Surveillance System Based on Deep Learning. 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE).
[4]. He, S., Shuai, Z., Zhou, Q., Bai, X., Cheng, M. M., & Zhang, J. (2020). An AI-based Crowd Monitoring System: Unseen Feature Learning and Context Reasoning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[5]. Wu, X., & Zhang, Z. (2019). A Survey on Learning to Detect Objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), 1936- 1959.
[6]. Yang, Y., Zhu, Y., Gao, L., Jiang, H., & Cao, X. (2020). A Survey on Object Detection in Optical Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 58(10), 7215-7238.
[7]. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2019). A Survey on Deep Transfer Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[8]. Yao, Q., Wang, D., Zhang, K., Chen, S., Liu, Q., & Gong, Y. (2019). Towards Making Unbiased Metric Learning: Adaptive Separation Loss. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
[9]. Wang, J., Ding, H., Zhou, Y., & Cheng, J. (2019). Vehicle Detection in Aerial Images: A Review and Benchmark Dataset. IEEE Transactions on Geoscience and Remote Sensing, 58(11), 7833-7852.
[10]. Zhang, H., Wang, Y., & Kong, F. (2020). Crowd Density Estimation via Adversarial Video Generation. 2020 IEEE International Conference on Multimedia and Expo (ICME).
[11]. Li, Z., & Zhang, Z. (2019). Crowd Counting with Deep Structured Scale Integration Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[12]. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards RealTime Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems (NeurIPS).
[13]. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic Understanding of Scenes through ADE20K Dataset. International Journal of Computer Vision, 127(3), 302-321.
[14]. Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2018). Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NeurIPS).
[15]. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., & Reed, S. (2016). SSD: Single Shot MultiBox Detector. European Conference on Computer Vision (ECCV)