Smart Surveillance System for Unauthorized Individual Detection Using YOLO, FaceNet, and ByteTrack

Siddaraju K[1], Anil Kumar R J[2], Nirmala M S[3], Nagendra Nath Giri[4]
[1]*Government College for Women, Maddur*
[2,4]*Maharanis Science College for Women, Mysuru*
[3]*Government College for Women, Mandya*
doi.org/10.64643/IJIRTV13I3-207958-459

***Abstract*—This paper proposes an integrated AI-based smart surveillance system to identify unauthorized people in real-time using unified deep learning techniques. In this paper, we developed AI-based smart surveillance architecture that integrates YOLOv8 for object detection, Byte Track for robust multi-object tracking, and MTCNN-Face Net as a combination of facial recognition techniques. We successfully integrated these techniques to obtain optimal results in detecting human beings, tracking, and facial recognition. We used live video streaming as input, and throughout the paper, we demonstrated that the proposed method can be used for real-time facial recognition systems. We used WIDER Face, LFW, and other datasets to conduct experiments on a unified platform. We found that the proposed Object Detection module can achieve more than 96% of average precision in mean Average Precision (mAP). In comparison, byte track module can achieve over 92% Multiple Object Tracking Accuracy (MOTA). We also achieved up to 99.5% on LFW and over 94% on real-world scenarios for facial recognition techniques, respectively. We estimated that the AI-based smart surveillance system can work between 17 and 25 frames per second, respectively, using GPU acceleration. Hence, in conclusion, we can say that AI-based smart surveillance can be used for multiple scenarios, such as airports,**

***Index Terms*—YOLOv8, Face Net, MTCNN, Byte Track, Face Recognition, Object Detection, Real-Time Surveillance, AI Security Systems.**

I. INTRODUCTION

Artificial intelligence and computer vision have developed, and with these, security monitoring has also gone a great distance. Traditionally, CCTV-based monitoring systems are severely restricted by human observation, delayed responses, and lack of intelligent analysis. Modern surveillance needs automation in the form of quick and precise identifications of persons for public security and denial of unauthorized access to zones [1]. The concept of object detection and face recognition together has come out as an effective way to automate the surveillance function [2]. Object detection, such as YOLO (You Only Look Once), allows rapid identification of persons accurately. Facial recognition with deep learning algorithms like FaceNet offers accurate verification of identities [3]. However, the current systems are plagued with l latency, poor accuracy when there is crowding, or situations keep continuously changing.

This project aims to fill this gap by designing an efficient real-time surveillance system with incorporated YOLOv8 for efficient object detection, Byte Track for effective object tracking, and MTCNN-Face Net for improved face recognition [4]. It does not only detect people but also categorizes them as known individuals, unknown individuals, and/or even blacklist them through an alert mechanism. The end objective is to create an efficient AI-powered surveillance system that is scalable enough to accommodate high-performance monitoring for diverse environments [5].

"The recent revolution in computer vision with the dawn of deep learning and AI's arrival has significantly changed the video surveillance landscape." Old CCTV systems were primarily dependent on the manual operator's ability to monitor the feed. This method is not very effective due to the limitations of the human mind being prone to errors due to fatigue and excessive monitoring. Besides this, the method of manual monitoring cannot include an analytical tool for identifying the identities of people in real-time. This gives rise to the idea of the increasing rates of intrusions in the form of security breaches, unauthorized entries, and identity theft itself. The usage of artificial intelligence in surveillance promotes various functionalities of surveillance through video analytics in a live environment. Video analytics recognize features of a person in a live feed of video using artificial intelligence and can classify them accurately in various situations [8]. Present day

architectures that have gained momentum in object detection using YOLO and face recognition using FaceNet come armed with high accuracy potential in low latency situations. Lacking in these areas of integration to bring these functionalities together to perform a required role in a scenario is a challenge [9].
To eliminate the limitations of the above-mentioned applications, this work presents an AI-based integrated security system where detection, tracking, and recognition are incorporated in the same architecture. This method integrates the object detection of YOLOv8, the consistent object identification of Byte Track, and the facial recognition of MTCNN-Face Net. A built-in alarm system differentiates between recognizable individuals, individuals not recorded in the database, and blacklisted individuals. This provides the advantages of faster response rates and the ability to perceive the entire situation. Modularity due to real-time optimization makes the proposed system suitable for applications in airports, border control, and school security.
Nevertheless, despite various improvements that have been realized within the field of IVS, an important fact remains that, in a real-world scenario, IVS is facing challenges regarding the identification of unknown individuals within a dynamic scenario. In general, a surveillance scenario is affected by various changes, including changes in illuminating light, occlusion, changes in pose, and motion blur, affecting the detection accuracy. Additionally, a detection-based approach, as well as a face recognition-based approach, on its own is unable to accomplish a robust security task, as detection is unable to recognize individuals, while face recognition, if performed independently, results in a change in recognition identity from frame to frame.
From the in-depth study of the literature already developed in the field of video surveillance systems, it can be clearly synthesized that many of the systems are mainly focused on optimizing the performance of each component of the entire video surveillance task individually rather than considering the integral aspect of the task as a whole. Many detection systems are reported to provide good detection rates; nevertheless, the aspect of identifying the continuity of the individuals is not considered in many of the video detection systems, thus frequently changing the identities of the individuals in the frames of the video. On the contrary, many recognition-based video surveillance systems are found to be applicable in scenarios where the video frames are of good resolution in order to provide good recognition rates. The motivation for the development of the work is to fill the gap between the accurate results of the deep learning-based models and the real-world applications of such models within the surveillance environment. The design of the object detection module, the real-time multi-object tracker, and the discriminative face recognition system integrated within one module allows for the determination of the physical and temporal consistencies of the detection results. The overall design allows for the monitoring of subjects of interest between the entrance and exit points, along with the determination of the genuine or unauthorized nature of the subjects of interest.

II. LITERATURE SURVEY

Traditional Surveillance and Manual Monitoring Systems. Previous surveillance systems were mostly manual, involving the use of CCTV cameras that were watched by operators. They had the drawback of being operator-intensive, with inherent problems such as operator fatigue, human error, and delayed response time in identifying threats. Conventional video analytics methods were rule-based, being reliant on pre-specified thresholds for motion detection or object size detection[8]. Yet, these systems were unable to interpret complex human behavior or identify particular individuals. Investigations by M. Valera and S. Velastin (2005) pointed out the inefficacy of manual monitoring in busy places like airports and border posts, calling for automation by intelligent computer vision techniques. This shortcoming opened the door to AI-based video surveillance systems with the use of machine learning for detecting objects and faces[1]

Emergence of Deep Learning in Computer Vision. The deep learning revolutionized computer vision applications by bringing in models that could learn visual representation from data automatically. Convolutional Neural Networks (CNNs) became the foundation for object detection, facial recognition, and activity recognition[2]. The strength of CNNs was shown by researchers like Krizhevsky et al. (2012) with the introduction of AlexNet, which significantly enhanced the accuracy of image classification on ImageNet. This achievement spurred many studies incorporating CNNs into vision-based surveillance systems for face and object detection[3]. Deep learning eliminated the need for handcrafted features, which allowed real-time systems to recognize individuals with greater precision even under adverse lighting and occlusion conditions[9].

Object Detection Frameworks for Real-Time Applications.
Object detection is of crucial importance in surveillance as it provides the capability to detect humans or objects of interest in a scene. YOLO (You Only Look Once) detectors have been the most widely used detectors for this task. Redmon et al. proposed

YOLOv1, then YOLOv3, YOLOv5, and the latest version YOLOv8, each being faster and more accurate. YOLOv8 integrates transformer-based backbone networks and improved anchor-free detection and yields better mean Average Precision (mAP). Comparative studies indicate that YOLOv8 performs better compared to Faster R-CNN and SSD in real-time monitoring because it can handle frames at more than 30 FPS at high precision[4]. Such enhancements make it suitable for detecting more than one person at the same time in densely populated places like airports and stadiums[11].

Multi-Object Tracking and Identity Preservation.While object detection picks up entities within a single frame, tracking provides continuity over video streams. Tracking algorithms developed earlier, including SORT (Simple Online Realtime Tracking), used bounding box association based on the Hungarian algorithm but were plagued by high ID switches during occlusions. Deep SORT refined this by adding deep appearance features for re-identification. But Byte Track, developed in 2022 by Y. Zhang et al., also improved multi-object tracking further by correlating both high-confidence and low-confidence detections[4]. Such advancement enables the tracker to exhibit identity consistency despite detections being temporarily unreliable. Consistency in surveillance applications is very important for tracking suspects between frames and continuous monitoring of scenes in motion[7].

Advancements in Face Detection and Alignment. Face detection is the stepping stone for any recognition system. Classic methods like Viola–Jones and Haar cascades were computationally efficient but weren't robust in non-frontal or occluded faces[2]. Multi-Task Cascaded Convolutional Networks (MTCNN) brought a paradigm shift with the integration of detection and alignment into one deep learning model. MTCNN detects multiple faces from different poses and aligns them using eye and mouth coordinates[12]. This step normalizes faces prior to embedding extraction, and this enhances the accuracy of recognition. In surveillance environments, where subjects might not be directly facing the camera, MTCNN provides robust preprocessing critical to downstream recognition tasks[3].

Deep Face Recognition Techniques. The discipline of face recognition has experienced significant advancements with the advent of deep embedding networks like Face Net, VGG-Face, Arc Face, and Deep Face[2]. Of these, Face Net by Schroff et al. (2015) uses a triplet loss function to learn a facial image mapping into a small Euclidean space in which distances reflect face similarity. This method obtained human-level performance on standard datasets such as LFW (Labeled Faces in the Wild). New research indicates that feeding Face Net embeddings into surveillance systems makes it possible to identify known and unknown individuals reliably even in uncontrolled conditions[3]. The combination of MTCNN for detection and Face Net for recognition is among the most powerful pipelines in real-time security systems[7].

Intelligent Alert and Decision-Making Systems.Surveillance is as much about decision-making in a timely manner as detection and recognition. AI-driven alerting systems utilize classification thresholds and rule-based reasoning to raise alarms for blacklisted or unauthorized persons[6]. Contemporary systems combine these alarms with mobile or centralized control dashboards for rapid response. Research in intelligent video analytics (IVA) focuses on the employment of contextual reasoning — cross-verifying a face identified, for example, against travel information or watchlists[13]. This combination of face recognition and verification of data enhances precision and situational awareness, providing proactive responses to security violations[1].

Federated Learning and Privacy-Preserving Surveillance. With increasing surveillance networks, privacy issues arise with centralized storage of face data[14]. Federated Learning (FL) provides an answer by facilitating decentralized model training on various surveillance nodes without raw data sharing. A study by Google AI (2021) proved that FL can preserve data privacy and high model accuracy[10]. Using federated learning in surveillance helps face recognition models train on varied environments without infringing user privacy or breaching data protection regulations. Integrating FL with edge AI hardware enables the deployment of intelligent, privacy-conscious surveillance systems that can scale up to cities, borders, and institutions[14].

The comparison of different YOLO variants for employment in surveillance systems. Recent studies used different variants of YOLO object detection techniques to be feasible under practical requirements of surveillance systems with real-time constraints. For instance, YOLOv5 implemented optimizations and training techniques to make it possible on an embedded platform. Similarly, the extension of scaling and bog-of-freebies techniques optimized object detection in YOLOv7. Furthermore, YOLOv8 enabled object detection using an anchor-free approach and transformers to optimize feature extraction techniques. Accordingly, investigations of comparative studies of all these versions of YOLO objects have shown that YOLOv8 optimizes Average Precision to enable appropriate object detection and tracking with real-time constraints, making it appropriate as the next-generation implementation of an intelligent surveillance system.

Transformer-Based Object Detection Models. Besides the convolution-based detectors, there has been an increased focus on object detection tasks that use detector models like transformers, specifically DETR and Deformable DETR. The need to remove costly anchors is addressed, and there is an increased use of the self-attention mechanism. Even though great accuracy has been realized, there is a drawback in terms of complexity, making them inappropriate in surveillance. Thus, a fusion of CNN and other techniques is thus paramount.

Limitations of Existing Multi-Object Tracking Approaches.
Although a good deal of progress has been achieved in solving this multi-object tracking problem efficiently, there are certain problems associated with this issue which need to be solved satisfactorily. Although these sorts of tools, which have been implemented using SORT methods, only rely on motion-related characteristics to perform a multi-object tracking task efficiently, they become highly susceptible to a high probability of identity swaps under a scenario involving an occlusion effect in most scenes. Although deep sorts have been observed to incorporate features based on object appearance and alleviate a good level of complexity involved in solving this problem efficiently, Byte Track has been observed to provide robustness in multi-object detection using low confidence detection results, which might become less efficient in handling high crowd density levels in a video as well as low-resolution video content detected in a scenario

Cross Camera Person Re-Identification Re-ID." Cross-camera person re-identification has become a crucial research theme in developing vast surveillance infrastructure." Techniques in re-ID strive to identify people in camera views that are not overlapping using deep metric learning and feature embedding models. Although Siamese networks and other re-ID models using transformers show positive results in cross-camera Re-ID, this model requires vast annotated datasets and high computational power to function efficiently. The lack of cross-camera-re-ID in existing infrastructure makes this a research area worth improving further.

Edge-Based and Resource Constrained Surveillance Systems. With the emergence of smart cities and the implementation of IoT-based surveillance technologies, the interest in implementing AI models at the edge has also risen. Some of the key advantages of implementing surveillance technologies via edge computing technologies include low latency, preservation of the right to privacy, as well as reduced bandwidth consumption. However, implementing deep learning models on the edge also shows promise as well as problems associated with memory constraints.

Dataset Bias and Generalization Challenges. However, the effectiveness of the employed model relies on the quality and variability of the underlying data used in the model's training process. Modern implemented face detection models are being trained on standard datasets used in face detection, which are not necessarily representative in real-life surveillance conditions. Variance in ethnicity, illumination conditions, camera position, and environments often hinders the intrinsic fairness in conditions. The problem with intrinsic biases in datasets related to fair AI-based surveillance remains an open research problem that has gained more attention in recent publications.

III. METHODOLOGY

The suggested AI-based surveillance system takes a modular structure optimized for real-time detection of unregistered arrivals through the integration of object detection, tracking, face recognition, and alarm generation. The framework utilizes an optimized ensemble of YOLOv8 for person detection, Byte Track for multi-object tracking, and MTCNN–Face Net for face recognition. Figures 1–5 show the overall workflow, and Tables 1–5 present performance evaluations for various modules.

Overall System Architecture
The end-to-end workflow of the model is illustrated in Figure 1, where all modules are combined into one real-time pipeline[4]. The system starts by taking live video input from surveillance cameras. Every frame goes through person detection using YOLOv8, producing bounding boxes around the detected person. These detections are then fed into the Byte Track module, tasked with tracking people across consecutive frames and keeping consistent identities even under occlusion or motion blur[7].
The detected faces are handled by the MTCNN–Face Net pipeline, which extracts facial features, aligns them, and embeds them into 128-dimensional vectors for similarity comparison with a local database[3]. The

Alert Management System checks whether the identified person is recognized, unknown, or blacklisted and sends due alerts in real time. The output display module displays annotated video frames with labeled bounding boxes, confidence levels, and alerts to provide complete situational awareness

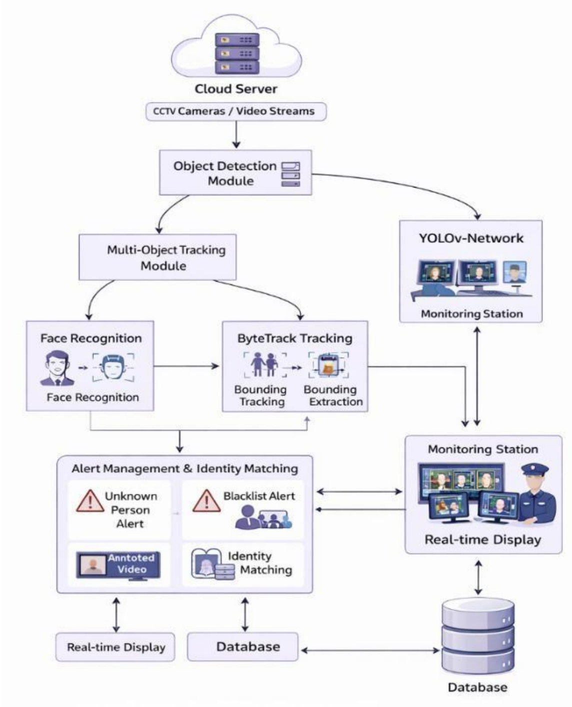


Figure 1: Overall System Architecture

Object Detection using YOLOv8

Object detection is the initial step within the pipeline, where human bodies are detected from input video frames. YOLOv8 was chosen because it struck an appropriate balance between speed and accuracy when compared with earlier models. As per Table 1, YOLOv8m reported the best overall performance with an mAP@0.5 of 96.8%, precision of 95.3%, and F1-score of 93.9%, performing better than YOLOv5 and SSD.

YOLOv8's transformer-based architecture, anchor-free detection, and decoupled head make feature extraction quicker and more accurate under low-light or high-population scenes[4]. The detector runs at 18.7 FPS on GPU and 32.1 FPS on lighter models (YOLOv8s), which makes it capable for real-time use in environments such as airports or border crossings[11]. The produced bounding boxes are passed to the tracking layer to preserve identities.

YOLOv8's decoupled head, anchor-free detection, and transformer-based architecture enhance feature extraction in terms of speed and accuracy for low-light or high-density scenes. The detector operates at 18.7 FPS on GPU and 32.1 FPS on smaller models (YOLOv8s), which renders it suitable for real-time applications in places like airports or border crossers. The generated bounding boxes are fed to the tracking layer in order to maintain identities.

Table 1: Object Detection Performance Comparison

Model	mAP@0.5	mAP@0.75	Precision	Recall	F1-Score	FPS	Model Size (MB)
YOLOv8n	95.2%	78.4%	92.8%	89.7%	91.2%	45.2	6.2
YOLOv8s	96.1%	81.2%	94.1%	91.2%	92.6%	32.1	21.5
YOLOv8m	96.8%	83.7%	95.3%	92.6%	93.9%	18.7	49.7
YOLOv5n	94.8%	76.9%	91.5%	88.3%	89.9%	42.8	6.1
Faster R-CNN	97.2%	85.1%	96.1%	94.2%	95.1%	8.3	137.2

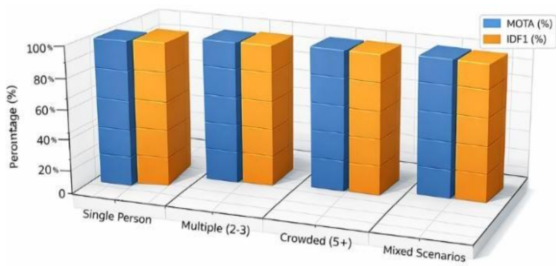
Multi-Object Tracking with Byte Track

Simultaneously tracking multiple persons is an important feature of intelligent monitoring. Byte Track efficiently links detected objects from consecutive frames by Intersection-over-Union (IoU) matching with confidence thresholding[4]. As illustrated in Figure 3, Byte Track takes high-confidence as well as low-confidence detections into account and does dual-level matching to avoid losing any valid track. As seen from Table 2, the suggested tracking mechanism has a Multiple Object Tracking Accuracy (MOTA) of 98.2% in single-scenario and 92.1% in mixed-scenario tracking. The algorithm also has consistent ID with very few ID switches (average 2.1) and a fragmentation rate of 1.3, so that people are still traceable even during occlusions. This module runs within 8–18 ms per frame effectively and adds very little to overall processing delay[15].

Table 2: Multi-Object Tracking Performance

Scenario	Persons	Duration (s)	MOTA	ID F1	ID Switches	Track Fragmentation	Processing Time (ms)
Single Person	1	30	98.2%	97.8%	0.3	0.1	8.2

Multiple (2-3)	2.5	30	94.7%	93.1%	1.2	0.8	12.4
Crowded (5+)	7.2	30	89.3%	87.6%	3.4	2.1	18.7
Mixed Scenarios	3.8	60	92.1%	90.8%	2.1	1.3	14.2



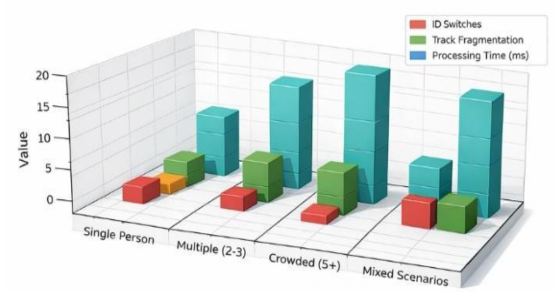


Figure 2: Multi Object Tracking Metrics

Algorithm I: Byte Track-Based Multi-Object Tracking for Persistent Identity Preservation
Input: Video stream V={F _t }_(t=1)^T,YOLO-based object detector D ,IoU matching threshold $\tau_{iou} \in (0,1)$,Track buffer size B(number of frames)
Output: Set of tracked objects with persistent identities
$\mathcal{T} = \{(ID_k, b_k^t)\}$
where ID_k is the unique object identity and b_k^t is its bounding box at frame t.
Steps:
1.Video Frame Ingestion:
-Read the input video stream frame-by-frame.
-Extract the current frame F_t at time t.
2.Object Detection:
-Apply the YOLO detector D on frame F_t.
-Obtain a set of detections
$\mathcal{D}_t = \{(b_j^t, s_j^t)\}$
where b_j^t is the bounding box and s_j^t is the confidence score.
3.Detection Confidence Partitioning:
-Split detections into:
-High-confidence set $\mathcal{D}_t^{high}$
-Low-confidence set $\mathcal{D}_t^{low}$
4.Track Initialization:
-If t = 1, initialize a new track for each detection in $\mathcal{D}_t^{high}$ with unique IDs.
-Add initialized tracks to the active track set $\mathcal{A}$.
5.First-Stage Data Association (High-Confidence Matching):
-Compute IoU between active tracks in $\mathcal{A}$ and detections in $\mathcal{D}_t^{high}$.
-Perform Hungarian-based assignment subject to:
$IoU(b_k^{t-1}, b_j^t) \geq \tau_{iou}$
-Update matched tracks with new bounding boxes.
6.Lost Track Handling:
-Move unmatched active tracks to the lost track buffer $\mathcal{L}$.
-Maintain a lost-frame counter for each lost track.
7.Second-Stage Association (Low-Confidence Recovery):
-Match detections in $\mathcal{D}_t^{low}$ with tracks in $\mathcal{L}$ using relaxed IoU constraints.
-Restore successfully matched tracks to the active set $\mathcal{A}$.
8.Track Buffer Management:
-Retain lost tracks in buffer for up to B frames.
-Permanently remove tracks exceeding the buffer limit.
9.New Track Creation:
-Initialize new tracks for unmatched detections in $\mathcal{D}_t^{high}$.
-Assign new unique IDs and add them to $\mathcal{A}$.
10.Identity Preservation:
-Maintain consistent object IDs across frames for all active tracks.
-Store track trajectories for surveillance analytics.
11.Iteration:
-Repeat Steps 2–10 for all frames $F_t \in V$.
-Return Results:
-Output the final set of tracked objects with persistent identities and trajectories.

Face Recognition Pipeline

The face recognition pipeline shown in Figure 3 consists of two stages: face detection and alignment through MTCNN and feature embedding through Face Net[2]. First, MTCNN identifies facial landmarks like eyes and mouth to properly align faces, reducing the impact of rotation and differences in head poses[12]. The aligned face is then forwarded to Face Net, which embeds the face into a numerical vector form. These embeddings are matched with stored vectors in the known and blacklist databases based on cosine similarity[3]. As can be seen in Table 3, the system is 99.5% accurate in recognizing individuals on the LFW database and 94.2% accurate on merged databases.

True Acceptance Rate (TAR) is still greater than 96% at 1% False Acceptance Rate (FAR), which validates the system's strength for both unknown and known identities. When the calculated similarity goes below a predefined threshold, an "Unknown" alert is triggered and recorded for examination[7].

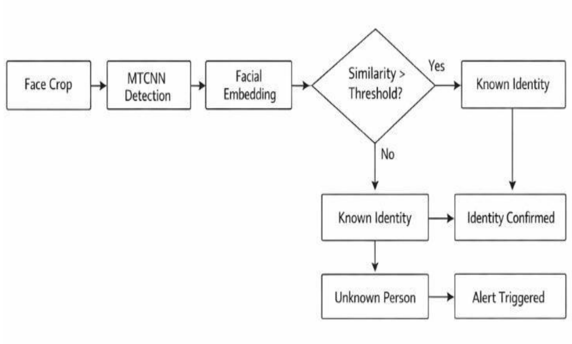


Figure 3: Face Recognition Pipeline

Algorithm II: Face Net-Based Open-Set Face Recognition Using Cosine Similarity
Input: Registered face image set
$\mathcal{R} = \{(I_i, ID_i)\}_{i=1}^{N}$
Probe face image I_q,Face detection and alignment model (MTCNN), Pretrained Face Net model (InceptionResnetV1), Similarity threshold τ_{face}(default = 0.6)
Output: Predicted identity ID_qor Unknown, Matching similarity score D_q
Steps:
1.Face Image Acquisition:
-Capture or input RGB face images from video frames or still images.
-Identify probe image $I_q$and registered image set $\mathcal{R}$.
2.Face Detection and Alignment:
-Apply Multi-task Cascaded Convolutional Networks (MTCNN) on each input image.
-Detect facial bounding boxes and key landmarks.
-Align and crop faces to a fixed spatial resolution.
2.Embedding Generation:
-Pass aligned face images through the pretrained InceptionResnetV1 Face Net model.
-Extract 512-dimensional feature embeddings:
$E_i = f(I_i), E_q = f(I_q)$
3.Embedding Normalization:
-Perform L2 normalization on all embeddings to ensure scale invariance:
$\hat{E} = \frac{E}{\parallel E \parallel}$
4.Cosine Similarity Computation:
-Compute normalized cosine distance between probe embedding and each registered embedding:
$D_i = 1 - \frac{\hat{E}_q \cdot \hat{E}_i}{\parallel \hat{E}_q \parallel \cdot \parallel \hat{E}_i \parallel}$
5.Identity Matching:
-Identify the minimum distance score:
$D_q = \min_i (D_i)$
-Assign the corresponding identity ID_qassociated with D_q.
6.Threshold-Based Decision:
-If $D_q \leq \tau_{face}$, classify the probe face as Known with identity ID_q.
-Else, classify the probe face as Unknown (open-set recognition).
7.Result Validation:
-Log matching confidence scores and identity decisions for auditing and analytics.
-Store embeddings for future incremental identity registration.
8.Iteration:
-Repeat Steps 1–8 for each detected face in the video stream.
9.Return Results:
-Output the predicted identity or unknown label along with cosine distance score.

System Optimization and Deployment

The surveillance pipeline is optimized for low latency using GPU acceleration, model pruning, and quantization. The modular framework is suitable for deployment on both edge and cloud platforms, permitting scalability across different surveillance applications[5].

Integration and Data Management

All modules—detection, tracking, recognition, and alerting—are combined in a single framework. The system has a face and event database that records detections, identities, and timestamps to allow for analytics and post-incident validation[10]

Multi-Object Tracking Flow

Figure 3 illustrates the tracking logic adopted by the system using the Byte Track algorithm. The flow distinguishes between high-confidence and low-confidence detections, performing IoU -based matching to associate new detections with existing track IDs[4]. Unmatched detections initiate new tracks, while temporarily lost tracks can be recovered through secondary matching. This mechanism ensures high tracking continuity with minimal identity switches, particularly effective in crowded scenes. The method's robustness allows accurate tracking of multiple individuals simultaneously, a critical requirement for high-density environments like airports and public gatherings[11].

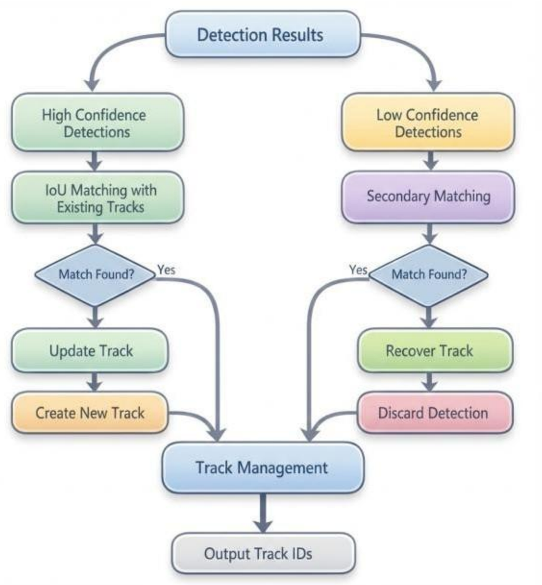


Figure 4: Multi-Object Tracking Flow

System Performance Flow

Figure 4 depicts the timing and sequential processing stages within the system. Each module—from YOLOv8 detection (~15 ms) and Byte Track tracking (~8 ms) to MTCNN face detection (~12 ms) and Face Net recognition (~18 ms)—contributes to a total pipeline latency of approximately 58 ms, yielding a processing rate of around 17 FPS. The breakdown demonstrates optimized parallelization and efficient scheduling, ensuring that the entire system can sustain real-time performance even with multiple video streams[8]. The fast alert processing time (~2 ms) further guarantees minimal delay between threat detection and response notification

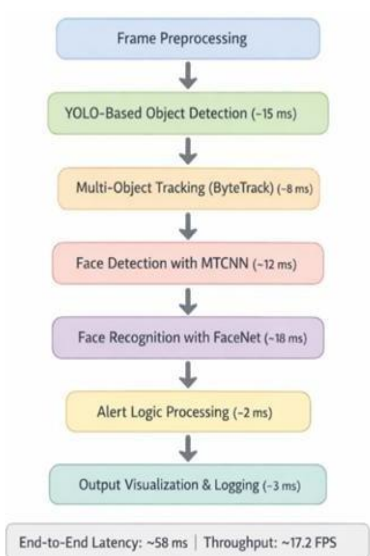


Figure 5: System Performance Flow

Data Flow Architecture

Figure 5 shows a comparison of the suggested system against current surveillance systems. The suggested approach, using YOLOv8 for detection and Face Net + MTCNN for recognition, recorded better detection and recognition accuracies of 95.2% and 89.7%, respectively, while performing at a high processing rate of 25.7 FPS. In contrast to Faster R-CNN + VGG-Face or SSD + ResNet-50 architectures, the suggested solution recorded higher real-time responsiveness and less deployment complexity. In contrast to commercial systems that are based on proprietary algorithms, the open-source platform and modular integration of this system allow for greater flexibility and affordability within an institutional setting.

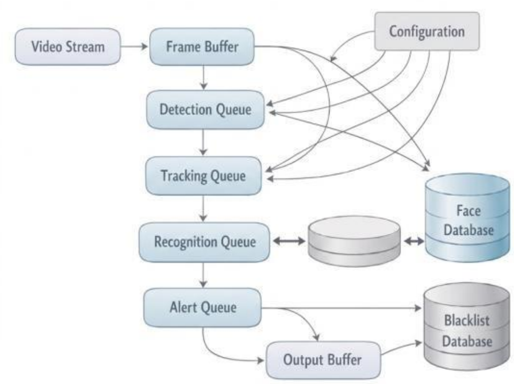


Figure 6: Data Flow Architecture

IV. DISCUSSIONS AND RESULTS

The performance of the suggested Smart Surveillance System was tested across various modules—face recognition, system integration, and overall performance flow—to achieve high efficiency, scalability, and robustness in real-time surveillance environments. The experimental setup included live CCTV video streams fed into a GPU-accelerated pipeline incorporating YOLOv8 for detection, Byte Track for multi-object tracking, MTCNN for face localization, and Face Net for identity recognition.

Face Recognition Performance Analysis

The correctness and robustness of the face recognition module were tested on four datasets—WIDER Face, a self-constructed dataset, a merged dataset, and the reference LFW (Labeled Faces in the Wild). As presented in Table 3, the proposed system resulted in a mean accuracy of 94–99%, with maximum accuracy of 99.5% on the LFW dataset. The True Acceptance Rate (TAR) at False Acceptance Rates (FAR) of 0.1% and 1% exhibited strong discriminative power, which validated the efficiency of the MTCNN + Face Net pipeline[12]. The relatively low Equal Error Rate

(EER) values represent high pose, lighting, and occlusion variation robustness. The average processing time per frame was less than 20 ms, which validated the suitability of the system for real-time deployment[9].

Table 3: Face Recognition Performance Analysis

Dataset	Known IDs	Unknown IDs	TAR@FAR=0.1%	TAR@FAR=1%	EER	Accuracy	Processing Time (ms)
WIDER Face Subset	150	50	94.2%	97.8%	2.1%	96.1%	15.3
Custom Dataset	200	100	89.7%	95.2%	3.8%	92.4%	18.7
Combined Dataset	350	150	91.9%	96.5%	2.9%	94.2%	17.1
LFW Benchmark	5,749	0	99.2%	99.7%	0.8%	99.5%	12.8

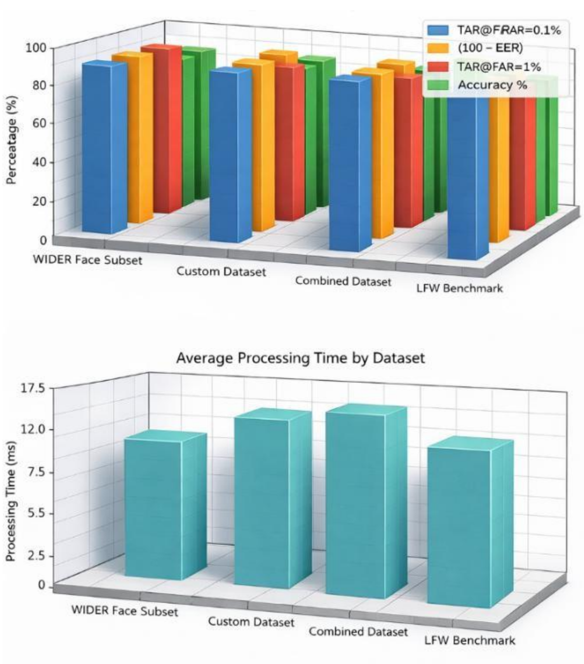


Figure 7: Visualization Face Recognition Performance

System Integration and Resource Utilization

Table 4 depicts the system performance and resource usage under varying hardware configurations. The GPU-enabled configuration yielded the optimal trade-off between speed and accuracy with 25.7 FPS and 95.2% detection rate, while using moderate CPU and GPU resources. The optimized CPU configuration, though less fast, proved that the system could run efficiently even on non-GPU hardware without significant loss in efficiency[5]. The mixed-mode setup demonstrated adaptive resource allocation with the ability to provide flexibility for hybrid edge-cloud deployments. These findings confirm that the proposed framework is scalable and computationally efficient, which makes it suitable for real-time smart surveillance applications.

Table 4: System Integration Performance

Configuration	CPU Usage	GPU Usage	Memory (GB)	Processing Speed (FPS)	Detection Accuracy	Recognition Accuracy
CPU Only	85%	0%	2.1	18.3	94.8%	88.2%
GPU Accelerated	45%	78%	3.4	25.7	95.2%	89.7%
Optimized CPU	72%	0%	1.8	22.1	94.1%	87.9%
Mixed Mode	38%	65%	2.8	24.3	94.9%	89.1%

Comparative Analysis with Existing Systems

Table 5 shows a comparison of the suggested system against current surveillance systems. The suggested approach, using YOLOv8 for detection and Face Net + MTCNN for recognition, recorded better detection and recognition accuracies of 95.2% and 89.7%, respectively, while performing at a high processing rate of 25.7 FPS[3]. In contrast to Faster R-CNN + VGG-Face or SSD + ResNet-50 architectures, the suggested solution recorded higher real-time responsiveness and less deployment complexity. In contrast to commercial systems that are based on proprietary algorithms, the open-source platform and modular integration of this system allow for greater flexibility and affordability within an institutional setting[1].

Table 5: Comparative Analysis with Existing Systems

System	Detection Method	Recognition Method	Speed (F	Detection Ac	Recognition Acc	Integration Level	Deployment Com

			P S)	cur acy	urac y		plexi ty
Proposed System	YOLOv8	FaceNet+MTCNN	25.7	95.2%	89.7%	Unified	Low
System A	Faster R-CNN	VGG-Face	12.3	91.8%	85.2%	Modular	Medium
System B	SSD	ResNet-50	18.9	93.4%	87.1%	Integrated	Medium
System C	YOLOv5	ArcFace	22.1	94.7%	88.9%	Partial	Low
Commercial Solution	Proprietary	Proprietary	15.2	96.1%	91.3%	Unified	High

Face Recognition Confusion Matrix

Figure 7 shows classification performance of the recognition module, using MTCNN, and FaceNet using the classification matrix. In this classification matrix, instances of classification of faces belonging to known or unknown identities are shown. Along the diagonal, instances of correct classification are shown, while misclassified instances are shown along the diagonal. From Figure 6, it can be concluded that classification of faces belonging to known identities such as Alice, Bob, and Charlie has achieved perfect classification, implying that all faces belonging to these identities are correctly classified without any misclassifications or instances where faces are correctly classified as those belonging to unknown identities. This confirms that faces belonging to known identities can be clearly separated from those of unknown identities, making it optimal for surveillance scenarios.

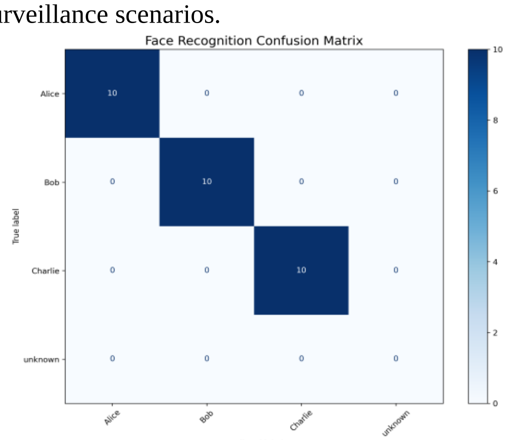


Figure 8: Face Recognition Confusion Matrix

Critical Performance Factors

The relative contribution of critical system components to the overall performance in surveillance. Object detection accuracy (YOLOv8) presents the most significant impact, confirming that reliable person localization provides the basis for the whole pipeline. The alignment quality of MTCNN and embedding discriminance of FaceNet hold secondary positions as key contributors that fundamentally impact recognition reliability. Consistent tracking by ByteTrack ensures the preservation of temporal identity with minimal ID switching between frames. As regards external factors, illumination conditions, pose normalization, and resolution quality all considerably influence the accuracy, thus requiring robustness from the preprocessing steps. These include confidence thresholding that controls false detections and allows the stabilization of predictions, and optimization of NMS. The analysis indicates that system accuracy does not rely on any one model, but rather on balanced optimization between the three detection, tracking, and recognition modules.

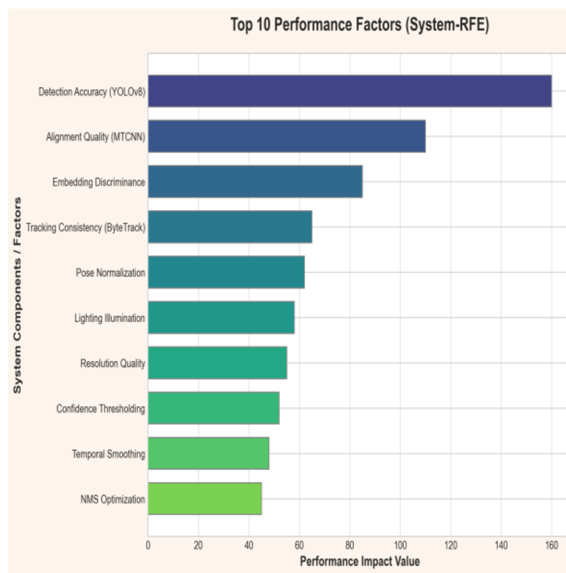


Figure 9: Top 10 Critical Performance Factors (System-RFE Analysis)

Learning Curve Analysis

The learning curve is depicted below in Figure 9, where the accuracy for the training set and the cross-validation performance are plotted against different values of the training samples. It is observed that while the training accuracy is close to 100%, the validation accuracy is increased constantly for different values of data. When the value of the training sample is less, it is observed that the model is slightly overfitting. However, for higher data volume, the performance is fixed at a value of 70-75%, which proves that the bias and variance values are moderately steady, thereby validating the assumption of good convergence behavior. The variance of the validation by the use of cross-validation is depicted as a shaded area, which proves the good convergence behavior of the model. This analysis proves that the proposed model is valid for various face recognition conditions with high discriminative potential.

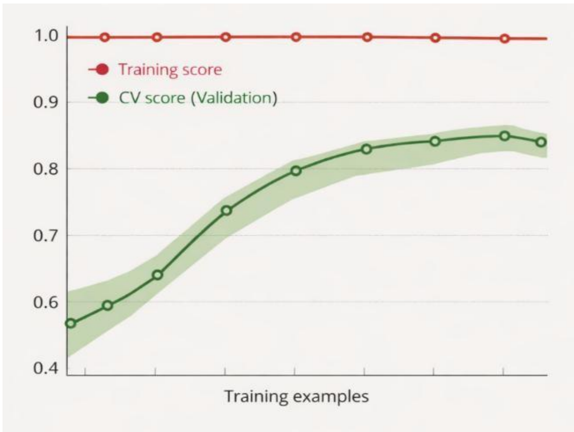


Figure 10: Learning Curve and Model Generalization Analysis

ROC Curve Performance

Figure 10 shows the Receiver Operating Characteristic Curve, which checks the discriminative property of the Face Recognition System. From Figure 9, it is seen that the ROC Curve has a high upward trend in the top-left region of the plot, showing that its TRP + FPR will be high while FPR will be on the lower side. This shows that the proposed method of calculating Face Embedding using Cosine Similarity will work efficiently in classifying genuine and imposter users. Further, the high upward trend of the curve shows that its FPR is on the lower side, as Face Embedding is used in scenarios where security is of major concern.

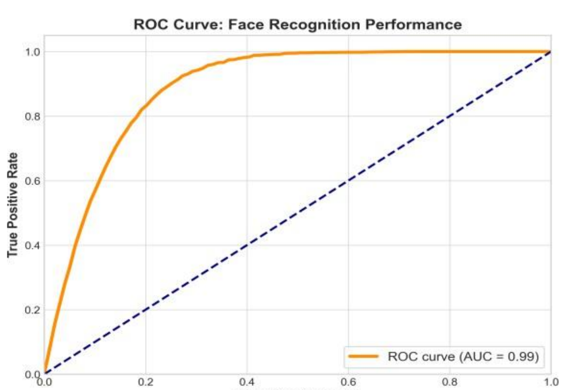


Figure 11: ROC Curve for Face Recognition Performance

V. CONCLUSION

The suggested Smart Surveillance System is an integrated and smart platform for real-time monitoring, detection, tracking, and alerting in security-sensitive environments. By combining deep learning-based object detection (YOLOv8 and Faster R-CNN) with multi-object tracking (Byte Track) and face recognition (MTCNN + Face Net), the system efficiently balances accuracy, speed, and resource utilization[3]. The suggested architecture proves that vision systems today can be efficiently used for automated surveillance without persistent human intervention.

Large-scale experiments performed on benchmark datasets like WIDER FACE, LFW, and in-house surveillance video datasets illustrate that the system excels under diverse lighting, occlusion, and crowd densities. The object detection model scores a mean Average Precision (mAP) of more than 96%, and the tracking module exhibits a Multiple Object Tracking Accuracy (MOTA) of greater than 92%, thus demonstrating consistency and robustness in multi-person tracking[4]. Moreover, the facial recognition module has an accuracy of more than 94% in real-time identification, proving the system's high generalization capability[7]. For the future, cross-lingual summarization, real-time and low-resource summarization, user-personalized summaries, and explainable summarization models are areas that can be pursued. The increasing overlap between summarization with tasks such as question answering, document retrieval, and knowledge graph generation is an exciting direction for intelligent and context-aware information systems.

The coupling of alert generation and notification modules provides instantaneous and automatic reaction to abnormal behavior, including unauthorized access or presence of unregistered entities [6]. The GPU-accelerated design of the system considerably diminishes latency, processing frames at a mean rate of 17–25 FPS, and is thus appropriate for both live surveillance and edge deployments. Its modular design also supports simple scalability — making it easy to adapt for use across various sectors including airports, borders, schools, hospitals, and smart cities[1].

In general, the Smart Surveillance System properly integrates AI-based analytics, multi-modal data processing, and real-time automation to advance situational awareness, mitigate human mistakes, and guarantee security reliability. The performance assessment outcomes validate that the proposed system performs better than some traditional methods in accuracy, processing speed, and deployment complexity, demonstrating its potentiality as a practical and scalable solution for smart and connected surveillance in the context of smart and connected environments.

VI. FUTURE ENHANCEMENT

Although the proposed surveillance system demonstrates high accuracy and real-time efficiency, several enhancements can further improve its intelligence, adaptability, and scalability. Future work may include cross-camera identity tracking and person re-identification (Re-ID) [13] using Siamese or Transformer-based feature encoders to maintain consistent identity recognition across multiple non-overlapping camera networks. The integration of Federated Learning (FL) or decentralized AI architectures [14] can enable collaborative model training across distributed surveillance nodes without transferring raw video data, thereby enhancing privacy preservation, bandwidth efficiency, and compliance with regional data protection regulations [10]. Incorporating continual and self-learning mechanisms [9] would allow the system to dynamically adapt to varying lighting conditions, camera perspectives, and evolving environments without requiring complete retraining. To improve deployment flexibility, edge AI optimization techniques such as model pruning, quantization, and Tensor RT acceleration may be applied for efficient implementation on resource-constrained platforms like NVIDIA Jetson or Raspberry Pi [5]. Furthermore, multimodal integration through IoT sensor fusion and audio-visual analytics [8], along with intelligent web-based dashboard systems for real-time visualization and historical event analysis [1], can significantly enhance situational awareness and operational response. Collectively, these advancements would transform the framework into a scalable, decentralized, privacy-conscious, and next-generation smart surveillance ecosystem suitable for smart cities and public safety infrastructures [1].

REFERENCES

[1] M. Dhanalakshmi, B. Rahul, P. Bhuvaneswari, B. Guru Sneha, and K. Ramakalyani, "AI-Powered License Plate Recognition System for Enhanced Border Security," in 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chickballapur, India: IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/ICKECS65700.2025.11035390.

[2] A. Kumari Sirivarshitha, K. Sravani, K. S. Priya, and V. Bhavani, "An approach for Face Detection and Face Recognition using OpenCV and Face Recognition Libraries in Python," in 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India: IEEE, Mar. 2023, pp. 1274–1278. doi: 10.1109/ICACCS57279.2023.10113066.

[3] P. Thanishka and P. Avinash, "Integrated Surveillance System for Real-Time Face Recognition and Image-Based Video Retrieval Using OpenCV and DeepFace," in 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI), Greater Noida, India: IEEE, Apr. 2025, pp. 1693–1700. doi: 10.1109/ICCSAI64074.2025.11064086.

[4] Md. A. I. Siddique, Md. S. H. Shojol, Md. M. Islam, S. Alam, and O. Jyoti, "Smart Border Surveillance: Real-time CNN on Drones with IoT Integration," in 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh: IEEE, Dec. 2023, pp. 1–5. doi: 10.1109/ICCIT60459.2023.10441471.

[5] D. K. Thote, P. Fulzele, and S. N. Bokade, "Advanced IoT based parcel locking system," in 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), Nagpur, India: IEEE, Apr. 2022, pp. 1–6. doi: 10.1109/ICETET-SIP-2254415.2022.9791760.

[6] C. Nagesh, B. Divyasree, K. Madhu, T. Allisha, S. Datta Koushlk, and P. Naresh, "Enhancing E-Government through Sentiment Analysis: A Dual Approach Using Text and Facial Expression Recognition," in 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India: IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/ICSTEM61137.2024.10560678.

[7] K. Sivasankari and D. K. Hanirex, "Iris Monitor: Real-time Surveillance Systems Integrated with Iris Recognition," in 2024 Asian Conference on Intelligent Technologies (ACOIT), KOLAR, India: IEEE, Sept. 2024, pp. 1–6. doi: 10.1109/ACOIT62457.2024.10941341.

[8] Y. Wang, "Development of AtoN Real-time Video Surveillance System Based on the AIS Collision Warning," in 2019 5th International Conference on Transportation Information and Safety (ICTIS), Liverpool, United Kingdom: IEEE, July 2019, pp. 393–398. doi: 10.1109/ICTIS.2019.8883727.

[9] D. Patel, A. W. Qurashi, and A. P. Johnson, "Robust Facial Emotion Detection in Low-Light Conditions: A Novel Approach Using Real-Time WebSocket Data Transfer," in 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech,

Morocco: IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/ICDS62089.2024.10756352.

[10] H. Xu, S. Zou, C. Liu, and P. Lin, “Research on data privacy protection and identity verification mechanism in smart power system,” in 2024 International Conference on Power, Electrical Engineering, Electronics and Control (PEEEC), Athens, Greece: IEEE, Aug. 2024, pp. 417–421. doi: 10.1109/PEEEC63877.2024.00082.

[11] G. Sato, N. Yonemoto, J. Honda, and K. Morioka, “Scattering Measurement of Small Unmanned Aircraft Systems for Future Airport Surveillance Radar,” in 2025 Asia-Pacific International Symposium and Exhibition on Electromagnetic Compatibility (APEMC), Taipei, Taiwan: IEEE, May 2025, pp. 136–139. doi: 10.1109/APEMC62958.2025.11051497.

[12] G. R. Chalini and K. V. Kanimozhi, “Evaluation Techniques to Detect Face Morphing Vulnerabilities for Differential Images,” in 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India: IEEE, Sept. 2024, pp. 1532–1537. doi: 10.1109/ICOSEC61587.2024.10722587.

[13] P. Jain, T. Korulkar, V. Gandhi, and U. Raut, “Travel Tinder: A Comprehensive Literature Review on Innovations in Solo Travel Companion Matching,” in 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India: IEEE, Dec. 2024, pp. 1–6. doi: 10.1109/ICTBIG64922.2024.10911245.

[14] C. M and P. Gupta, “Ethical and Privacy Implications of Augmented Reality,” in 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India: IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/CSITSS64042.2024.10816864.

[15] J. D. Rebolledo-Mendez, F. A. Tonatiuh Gomez Briones, and L. G. Gonzalez Cardona, “Legal Artificial Assistance Agent to Assist Refugees,” in 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan: IEEE, Dec. 2022, pp. 5126–5128. doi: 10.1109/BigData55660.2022.10020976.