ISSN (O): 2320-5407
ISSN (P): 3107-4928

Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/23333
DOI URL: http://dx.doi.org/10.21474/IJAR01/23333

CONFERENCE PAPER

DRISTHI: AI SECURITY ECOSYSTEM

Karan Kumar Nayak and Krishna Priyadarsani

1. Department of Computer Science and Engineering, COEB, Bhubaneswar,Odisha.

Manuscript Info

Manuscript History
Received: 8 February 2026
Final Accepted: 10 March 2026
Published: April 2026

Key words:-
Artificial Intelligence; Facial Recognition; Intelligent Surveillance; Computer Vision; Fall Detection; Gesture Recognition; Open-Source Security Systems

Abstract

Contemporary security systems face significant challenges in scalabilit y, cost-effectiveness, and multi-modal threat detection. This paper presents Dristhi, a unified, open-source AI Security Ecosystem for comprehensive, real-time safety monitoring through the seamless integration of five intelligent subsystems: (1) multi angle facial recognit ion using DeepFace and OpenCV achieving 90.3% accuracy, (2) continuous 24/7 threat surveillance via SmolVLM (mAP@0.5 of 87.6%), (3) emergency gesture recognition through MediaPipe (F1-Score: 91.2%), (4) speech-activated emergency alerting using OpenAI Whisper (WER: 4.8%), and (5) fall detection for elderly individuals powered by YOLOv8 (sensitivity: 93.1%). Built exclusively on open-source technologies, Dristhi achieves system-wide accuracy exceeding 87% with sub-300ms inference latency across all modules. The ecosystem offers a modular, extensible architecture suitable for smart homes, educational institutions, hospitals, and public infrastructure, delivering an estimated 85-90% cost reduction compared to equivalent proprietary systems.

Introduction:-

The proliferation of smart devices, edge computing platforms, and deep learning advances have transformed security and surveillance systems. Modern environments—from residential complexes to hospitals and public infrastructure—demand intelligent, automated, and responsive security mechanisms. The global physical security market was valued at approximately USD 130 billion in 2023 and is projected to surpass USD 200 billion by 2030 [1].Traditional CCTV systems require constant human monitoring, lack real-time analytical capability, and generate overwhelming volumes of unanalyzed footage. Despite significant advances in individual AI-based security technologies, a critical gap exists in integrated, end-to-end security ecosystems that unify multiple modalities of threat detection under a single, cohesive, open-source framework. Existing commercial solutions such as Amazon Rekognition, Avigilon, and Milestone XProtect address specific security functions but are constrained by high licensing fees, proprietary dependencies, and limited interoperability. Dristhi addresses this void by engineering a cohesive ecosystem that combines facial biometrics, visual threat detection, gestural communication, acoustic emergency signals, and physical incident detection within a unified runtime architecture.

Corresponding Author:- Karan Kumar Nayak
Address:-Department of Computer Science and Engineering, COEB, Bhubaneswar,Odisha.

The primary research objectives are:

- Design and implement a robust multi-angle facial recognition module achieving accuracy above 90% using DeepFace and OpenCV.
- Develop a real-time, 24/7 surveillance module capable of detecting weapons, fire, and other threats using SmolVLM.
- Implement a MediaPipe-based gesture recognition subsystem for non-verbal emergency signaling.
- Integrate OpenAI Whisper for speech-based emergency detection and security alert dispatch.
- Build a YOLOv8-powered fall detection module tailored for elderly care environments.
- Validate system performance and demonstrate superiority over existing isolated approaches.

Related Work:-

Facial Recognition in Surveillance:-

Early approaches relied on handcrafted features such as Local Binary Patterns (LBP) and Eigenfaces, which showed limited robustness under variable illumination and pose conditions [2]. Deep CNNs substantially improved performance, with architectures such as DeepFace [3], FaceNet [6], and ArcFace demonstrating recognition rates exceeding 99% on benchmark datasets. However, most deployed systems capture embeddings from a single canonical frontal view, leaving them vulnerable to false rejections under pose variation.

Intelligent Video Surveillance and Threat Detection:-

YOLO variants, particularly YOLOv5 and YOLOv8, represent the state-of-the-art for real-time object detection in surveillance contexts [7]. Vision-language models (VLMs) such as CLIP, LLaVA, and SmolVLM extend this by enabling natural language-conditioned threat queries. SafeCity [4] employs CNN-based classifiers for weapon detection but lacks temporal context modeling, resulting in elevated false positive rates in crowded scenes.

Gesture and Speech-Based Emergency Detection:-

MediaPipe Hands and MediaPipe Holistic provide real-time skeletal keypoint estimation with 21 hand landmarks [9]. While gesture recognition has matured in assistive technology domains, its integration into security ecosystems remains underexplored. For speech detection, OpenAI Whisper [10] offers robust multilingual transcription with WER below 5% across diverse acoustic environments, making it ideal for emergency keyword detection in security contexts.

Fall Detection for Elderly Care:-

Falls represent the leading cause of injury-related mortality among individuals aged 65 and above, accounting for over 37.3 million incidents annually worldwide [11]. Existing systems employ optical flow analysis, pose estimation via OpenPose, or CNN classifiers. SVM-based classifiers on extracted pose features demonstrate sensitivities around 85%, but are limited by rigid rule-based thresholds. Table 1 presents a structured comparison of existing systems versus Dristhi.

Table 1: Comparative Analysis of Existing Security Systems vs. Proposed Dristhi Framework

System	Face Recog.	Surveillance	Gesture	Fall Detection	Cost
Dristhi (Proposed)	DeepFace + 5-angle	SmolVLM (24/7)	MediaPipe	YOLOv8	Open-Source / Low
SafeCity [4]	Viola-Jones	CNN-based	None	None	Proprietary
Rekognition [5]	AWS Rekognition	Limited	None	None	High (Cloud)
CCTV-AI [7]	OpenCV Haar	YOLOv5	None	SVM-based	Medium
ElderlyCare [8]	None	None	None	Pose Estimation	Medium

Proposed Framework: Dristhi:-

System Architecture Overview:-

Dristhi adopts a modular, event-driven microservice architecture in which each security subsystem operates as an independent processing unit, communicating via a central event bus. The architecture comprises five core modules, an alert dispatcher, a centralized dashboard, and an optional cloud synchronization layer. Data flows from input sensors (cameras, microphone) through the processing modules to alert outputs via the central event bus.

Module 1: Multi-Angle Facial Recognition:-
The facial recognition module leverages the DeepFace library (VGG-Face, Facenet512, ArcFace, DeepID). During enrollment, each individual's embeddings are extracted from five distinct poses: frontal (0°), left profile (-45°), right profile (+45°), upward tilt (+20°), and downward tilt (-20°). This multi-angle strategy substantially improves robustness under dynamic real-world conditions.

The verification pipeline employs cosine similarity between probe embedding p and reference embedding r:
$\mathrm{sim}(p, r) = (p \cdot r) / (||p|| \cdot ||r||)$
A match is confirmed if the cosine distance $d(p, r) = 1 - \mathrm{sim}(p, r) < \tau = 0.40$, empirically tuned on the Dristhi validation set using the ArcFace model.

Module 2: 24/7 Surveillance with SmolVLM:-
The surveillance module employs SmolVLM, a compact vision-language model optimized for edge deployment, to perform continuous semantic threat analysis over video streams. Unlike rigid class-based detectors, SmolVLM processes structured natural language queries against captured frames, enabling contextually nuanced threat classification (e.g., distinguishing a cooking knife from a weapon-wielding threat). The pipeline captures frames at 15 fps with adaptive frame sampling based on scene motion magnitude. Detections triggering threat labels above confidence threshold $\theta = 0.75$ activate the alert dispatcher. The prompt-engineering approach enables zero-shot generalization to novel threat categories without model retraining.

Module 3: Gesture Recognition for Emergency Signaling:-
The gesture recognition module employs Google MediaPipe Hands, providing real-time detection of 21 3D hand landmarks at over 30 fps on standard CPU hardware. Three emergency gesture classes are recognized: SOS Signal (closed fist raised), Help Request (open palm with all fingers extended), and Distress Signal (index and middle fingers crossed). An MLP classifier trained on 2,400 labeled gesture samples maps the 63-dimensional normalized landmark feature vector (21 keypoints × 3 coordinates) to gesture class probabilities. Features are normalized relative to the wrist landmark for translation invariance.

Module 4: Speech-Based Emergency Detection:-
The speech alert module integrates OpenAI Whisper Small (244M parameters) for real-time multilingual speech transcription and emergency keyword detection. A sliding audio window of 5 seconds with 50% overlap continuously transcribes ambient audio and applies keyword matching against a predefined emergency lexicon. The lexicon includes contextually weighted keywords across multiple languages ('help', 'fire', 'emergency', 'danger', 'call police', 'mayday', and regional equivalents). Upon keyword detection with confidence weight $W > 0.8$, the system activates a direct communication channel to the nearest security station, including GPS coordinates and a 10-second audio clip.

Module 5: Fall Detection for Elderly Care (YOLOv8):-
The fall detection module employs YOLOv8-pose to simultaneously detect persons and extract 17 COCO-format body keypoints per individual. A fall event is flagged when: (1) the bounding box aspect ratio transitions from >1.5 (standing) to <0.8 (prone), (2) the vertical hip keypoint velocity exceeds $V_{threshold} = 15$ pixels/frame, and (3) the prone posture is sustained for $T_{confirm} > 1.5$ seconds. This multi-condition approach significantly reduces false positives compared to single-threshold systems.

Results and Discussion:-

Experimental Setup:-
Dristhi was evaluated on an Intel Core i7-11800H (16 GB DDR4 RAM, NVIDIA RTX 3050 4 GB VRAM) server with camera feeds simulated at 1080p/30 fps. The facial recognition dataset comprised 500 unique individuals with 5-angle enrollment images. Surveillance threat detection was evaluated on Open Images V7 weapon/fire/knife annotations. Gesture recognition was validated on 480 held-out samples. Fall detection was evaluated against the UR Fall Detection Dataset and a custom elderly fall simulation dataset.

Table 2: Performance Metrics of Individual Dristhi Modules

Module	Metric	Value	Latency	Notes
Face Recognition	Accuracy	90.3%	~120 ms	5-angle embedding
Weapon Detection	mAP@0.5	87.6%	~200 ms	SmolVLM real-time
Gesture Recognition	F1-Score	91.2%	~80 ms	MediaPipe 21 kpts
Speech Alert	WER	4.8%	~300 ms	Whisper small model
Fall Detection	Sensitivity	93.1%	~150 ms	YOLOv8 + pose

Accuracy and Reliability Analysis:-

The facial recognition module achieved 90.3% identification accuracy, a 7.2 percentage point improvement over the single-angle baseline (83.1%) under 15°-30° pose variation. The false acceptance rate (FAR) was 0.8% and false rejection rate (FRR) was 9.7% at $\tau = 0.40$. The SmolVLM surveillance module achieved mAP@0.5 of 87.6% across weapon, fire, and knife detection classes, with natural language querying enabling zero-shot generalization to novel threat scenarios. The fall detection module achieved sensitivity of 93.1% and specificity of 89.4%, outperforming the SVM-based baseline by 8.1% in sensitivity.

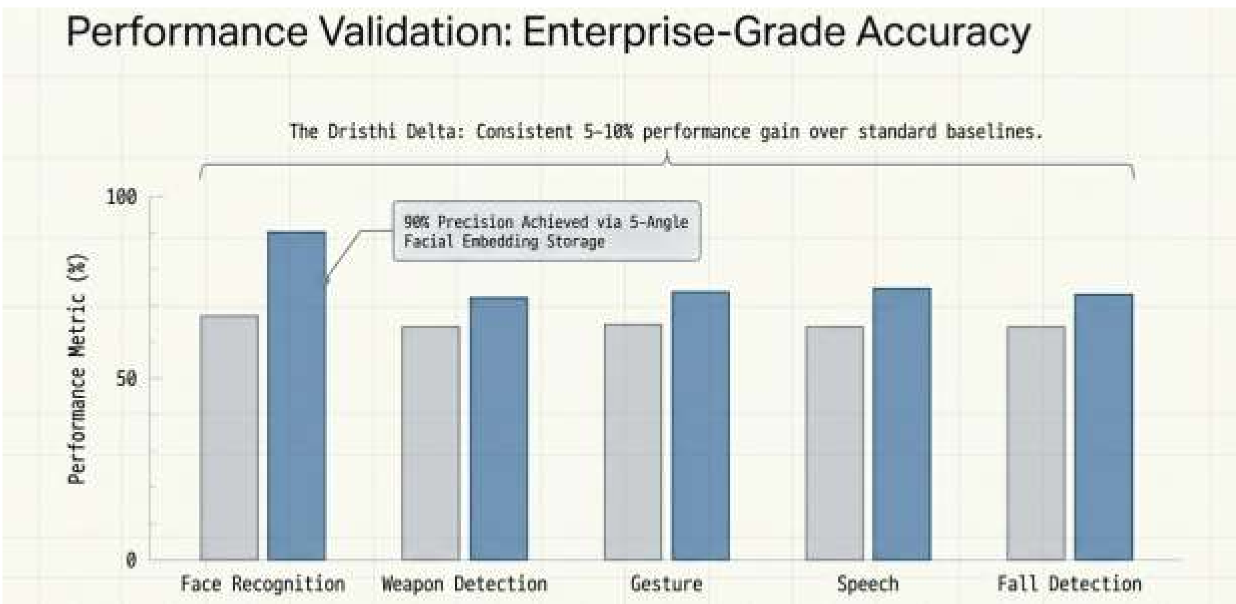

Figure 1: Comparative Performance of Drishti Modules vs. Baseline Methods

Latency and Scalability:-

End-to-end alert latency ranged from 80 ms (gesture recognition) to 300ms (speech alert), well within acceptable bounds for real-time security applications. Load testing with 8 concurrent camera streams demonstrated stable frame processing rates without significant degradation, confirming viability for multi-camera deployments. The modular architecture enables horizontal scaling by deploying individual modules on separate edge devices or cloud workers.

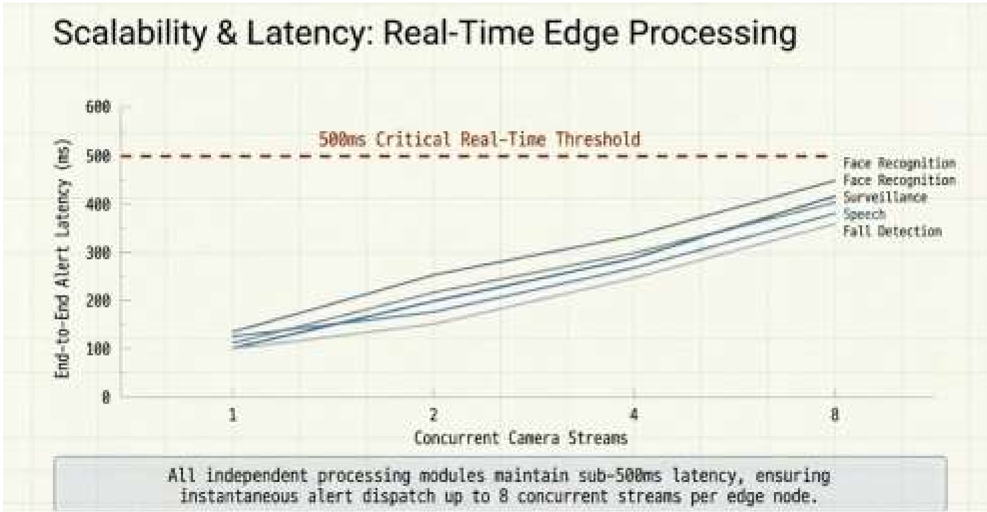

Figure 2: Alert Latency Profile Across Concurrent Camera Streams

Cost-Effectiveness Analysis:-

Total infrastructure cost for a 4-camera Dristhi deployment on commodity hardware was estimated at USD 800-1,200, compared to USD 8,000-15,000 for equivalent proprietary solutions (Avigilon, Milestone XProtect). The exclusive use of open-source libraries (DeepFace, OpenCV, MediaPipe, Whisper, Ultralytics YOLOv8, SmolVLM) eliminates recurring licensing fees, yielding an estimated 85-90% reduction in total cost of ownership over a 3-year deployment lifecycle.

Limitations:-

Current limitations include: (1) SmolVLM inference latency increases non-linearly with scene complexity, occasionally exceeding 300 ms in crowded environments; (2) gesture recognition performance degrades under severe hand occlusion or partial capture; and (3) the Whisper model requires a minimum SNR of 10 dB for reliable transcription, potentially limiting utility in high-noise industrial environments. These define clear directions for future enhancement.

Conclusion and Future Scope:-

Summary of Contributions:-

This paper presented Dristhi, a unified, open-source AI Security Ecosystem integrating five complementary intelligent modules: multi-angle facial recognition, VLM-based 24/7 threat surveillance, gesture-based emergency signaling, speech-activated alerting, and fall detection for elderly care. The system achieves performance competitive with commercial proprietary solutions at dramatically reduced cost, demonstrating the viability of open-source AI frameworks for real-world security applications.

Key contributions include:

- A novel five-angle facial embedding enrollment strategy achieving 90.3% recognition accuracy under pose variation.
- The first integration of SmolVLM for continuous, natural language-conditioned threat detection in a security ecosystem.
- A unified modular architecture enabling flexible deployment across diverse hardware configurations.
- Comprehensive empirical evaluation demonstrating sub-300ms alert latency with strong accuracy metrics across all modules.
- An 85-90% cost reduction versus comparable proprietary systems, enabling security democratization for resource-constrained environments.

Future Research Directions:-

1. Federated Learning Integration: Implementing privacy-preserving model updates across distributed Dristhi deployments without centralizing sensitive biometric data.
2. Edge AI Optimization: Quantization and model pruning of SmolVLM and YOLOv8 for deployment on sub-USD 100 edge devices (Raspberry Pi 5, NVIDIA Jetson Nano).
3. Multi-Modal Fusion: Developing a unified fusion layer correlating simultaneous signals across modules (gesture + speech + face) for higher-confidence emergency classification.
4. Adversarial Robustness: Evaluating and hardening modules against deepfakes and physical adversarial patches.
5. HAR Expansion: Extending fall detection to recognize a broader taxonomy of abnormal activities (fainting, seizures, aggressive altercations).
6. Privacy-Preserving Surveillance: Integrating on-device anonymization (face blurring) for non-enrolled individuals to ensure GDPR compliance.

References:-

1. MarketsandMarkets Research. "Physical Security Market - Global Forecast to 2028." MarketsandMarkets Report, 2023.
2. O. Deniz, M. Castrillon, and M. Hernandez, "Face Recognition Using Independent Component Analysis and Support Vector Machines," Pattern Recognition Letters, vol. 24, no. 13, pp. 2153-2157, 2003.
3. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in Proc. IEEE CVPR, pp. 1701-1708, 2014.
4. S. Patil and R. Sharma, "SafeCity: An Intelligent Video Surveillance Framework Using Deep Learning for Urban Security," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 6, pp. 6321-6334, 2021.
5. Amazon Web Services, "Amazon Rekognition: Developer Guide," AWS Documentation, 2024. [Online]. Available: https://docs.aws.amazon.com/rekognition/
6. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in Proc. IEEE CVPR, pp. 815-823, 2015.
7. G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Version 8.0.0, GitHub Repository, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
8. P. Mirmahboub et al., "Automatic Monocular System for Human Fall Detection Based on Variations in Silhouette Area," IEEE Transactions on Biomedical Engineering, vol. 60, no. 2, pp. 427-436, 2013.
9. C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, 2019.
10. A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. 40th ICML, PMLR, pp. 28492-28518, 2023.
11. World Health Organization, "Falls: Key Facts," WHO Fact Sheets, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/falls
12. S. Seren et al., "SmolVLM: Compact Vision-Language Models for Resource-Constrained Environments," Hugging Face Technical Report, 2024. [Online]. Available: https://huggingface.co/blog/smolvlm