International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I12P109

Volume 73 Issue 12, 95-123, December 2025
© 2025 Seventh Sense Research Group®

Original Article

Hybrid Deep Visual Intelligence Framework for Public Health and Safety Enforcement

R. Gayathri[1], Tan Kuan Tak[2], B. Sivaneasan[3], Siva Shankar S[4]

[1]*Department of Electronics and Telecommunication Engineering,*
JSPM's Bhavarabai Sawant Institute of Technology and Research, Pune, Maharashtra, India.
[2,3]*Engineering Cluster, Singapore Institute of Technology, Singapore.*
[4]*Department of Computer Science and Engineering, KG College of Engineering and Technology, Hyderabad, India.*

[1]*Corresponding Author : drgayathriradhakrishnan@gmail.com*

Received: 27 September 2025 Revised: 18 November 2025 Accepted: 25 November 2025 Published: 19 December 2025

Abstract - *This study highlights the urgent need for enhanced visual monitoring technologies to support public health and safety enforcement in crowded situations, in which individuals could potentially violate compliance and other norms. Most traditional surveillance methodologies lack either the scalability or precision to squash complaints about compliance violations – for example, improper mask usage, failure to maintain a minimum social distance, or other visible anomalies. To address these issues, we designed a Hybrid Deep Visual Intelligence Framework that integrates multiscale convolutional neural networks (e.g., ResNet-101, YOLOv8) for object detection with a sequence of 3D CNNs for temporal activity modeling. We also utilized Vision Transformers with cross-attention that spans external contextual metadata. The input data source was multimodal (including closed-circuit television footage, drone footage, and thermal imagery) to derive compliance scores or alerts in real-time. The evaluation for the constrained performance refers to independent public datasets representing four scenarios - indoor hospitals, outdoor markets, nighttime transport hubs, and large event stadiums using UCF-Crime datasets, Okutama-Action datasets, and AI City Challenge for autonomous vehicle innovative technology. The performance results demonstrated superior performance, achieving up to 96.4% detection accuracy and anomaly detections with F1-scores of 0.89, compared to traditional approaches, which were improved by 5-10%. The results included significant processing frame rates of greater than 24 frames per second that facilitated near-real-time execution. In conclusion, our proposed solutions represent a valid, privacy-compliant deployment option that adapts and scales for enforcement purposes, demonstrating significant options for supporting public health compliance and security in complex public spaces.*

Keywords - *Deep Learning, Video Surveillance, Public Health Compliance, Anomaly Detection, Multimodal Fusion, Real-Time Monitoring.*

1. Introduction

In recent years, the rapid expansion of advanced sensing and recording technologies has resulted in the unprecedented expansion of the amount of image and video data generated in public spaces (Rajitha, 2021). Sensing and recording devices such as surveillance cameras, drones, body-worn sensors, and thermal imaging cameras are now increasingly used to monitor urban conditions, transportation and logistics hubs, healthcare facilities, and gatherings of people in a variety of public spaces (Lee et al., 2023). These technologies were quickly adopted due to their perceived utility for enhancing security and operational efficiency. However, in the case of emerging public health threats to compliance with pandemic-related measures, their ability to uphold public health could also be very relevant - and very critical - to public sector services and public health institutions (Haley et al., 2025). Physical distancing, mask usage, and limited occupancies were critical in curbing disease spread in crowded public spaces, but monitoring compliance manually in large-scale scenarios with large numbers of people and places is resource-intensive, often results in errors due to human observation mistakes, and is often impractical if conducted in real time, thus leading to a counterproductive approach to managing public health. Intended or not, this is precisely why operational intelligence, and by extension, automated systems designed to analyze large and vast streams of visual data intelligently, are necessary and important (Qaraqe et al., 2024). Conventional video surveillance and pandemic enforcement have traditionally relied upon rule-based systems, or classical machine learning methods (Ardabili et al., 2023; Vashishth et al., 2024). Such conventional approaches typically employ hand-constructed features and rely heavily on fixed acceptance regions, which are not easily adapted to the complex dynamics of the real world, such as changing lighting, viewing perspectives, or crowd densities. Furthermore, classic systems excel at modeling fixed spatial characteristics but struggle to model the temporal characteristics of video, resulting in overall systems that perform poorly in tasks such as anomaly detection and activity recognition.

As a result, considerable research has been conducted to evaluate deep-learning approaches, which offer better generalization and ease of use when scaling. First, for object detection and segmentation tasks, Convolutional Neural Networks (CNNs) (Thirunagari et al., 2023) provide significantly improved accuracy over traditional methods. Furthermore, for modeling extended temporal sequences, recurrent and 3D convolutional models have proven successful for applications where video comprises a significant portion of the input to a machine learning model. Finally, attention-based architectures (e.g., Vision Transformers) have opened new opportunities for capturing long-distance dependencies and contextual relationships when modeling spatial-temporal aspects of visual data.

Existing research has employed various deep learning systems to address different aspects of public safety monitoring. For instance, YOL (Mostafa et al., 2024) and Faster R-CNN (Sahu et al., 2024) models have been used for real-time person detection and mask detection in images, while various 3D CNNs (Natha et al., 2025) and LSTM (Long Short-Term Memory) networks have been adopted for activity recognition and behavior classification from video footage. Despite these individual successes, past systems have been created independently, limiting their utility in a real-world compliance monitoring context, which would benefit from both fine-grained object detection and analysis of patterns of behavior over time (Arifuzzaman et al., 2024). Similarly, more traditional solutions ignore the value of multimodal data fusion: using potential alternative sources of information, such as thermal imagery, depth data, and environmental metadata, would allow for improved detection in challenging conditions, such as low light, occlusion, or crowding.

While many research contributions report high accuracy in well-defined experimental conditions, only a few have addressed aspects related to real-world outcomes, such as processing time, software scalability, privacy safeguards, and implementation features that support continuous learning. When a user validates an alert in real time, depending on how the enforcement system is set up, organisations may want to process data at an appropriate frame rate or delays for a timely notification, without overloading the computing infrastructure or affecting user privacy. Additionally, if the representative features of a scene evolve, or other compliance scenarios change, such as new legislation, then the user needs adaptive features to retain performance. To address these knowledge gaps and research challenges, this study proposes a Hybrid Deep Visual Intelligence Framework for image and video systems in the context of public health and safety enforcement. This framework aims to offer spatial imagery feature extraction at multiple scales, sequential activity modelling, and cross-modal fusion as a cohesive architecture capable of functioning across different environmental settings. This automation framework utilizes ResNet-101 and YOLOv8 for object and mask detection, 3D CNNs for capturing spatiotemporal behaviors in the data streams, and Vision Transformers with cross-attention layers that enable the addition of metadata, such as time of day and geographic location. The combination of CCTV camera footage, drone video, and thermal video also affords multi-modal data streams, and integrated data pipelines can support associated data input and output, and accountable compliance scores.

The method introduced is tested on a wide variety of publicly available data-sets, including UCF-Crime, Okutama-Action and the datasets from the AI City Challenge, in order to get an assessment of performance in contextually relevant situations: (1) Indoor Hospital Surveillances; (2) Outdoor, Crowded Markets; (3) Nighttime Transportation Hubs; and (4) Large Scale Events in stadiums.

Experimental results demonstrate that the approach proposed is superior to more standard CNN and transformer baselines in detection accuracy, anomaly detection F1 score, compliance scoring precision at real-time speeds, and these results point to the great potential for integrated deep learning frameworks to enable scalable, privacy-compliant public health enforcement systems adaptable to changing operational environments.

1.1. Research Question and Problem Statement

Building on the potential shortcomings and issues with existing surveillance solutions for anomaly detection involving public health compliance and safety, this study encompasses the research question: How can a hybrid deep learning framework incorporating multiscale spatial detection, temporal modeling, and multimodal data fusion be developed to monitor and frame violations of public health compliance measures in real-time in various settings? The fundamental problem will be the ability to develop a framework that provides a high level of detection accuracy and operates efficiently, but also adaptively in variable lighting conditions, crowds, and camera views. The main objectives of the study are as follows.

- To create and utilize a hybrid deep learning architecture that can integrate spatial, temporal, and contextual intelligence for public health compliance action monitoring in real-time.
- To compare the viability of the proposed framework and the baseline state-of-the-art methods on multiple datasets and disparate scenarios with respect to detection performance, infringement behaviour recognition performance, and latency speed.

- To create an innovative architectural framework that is legally compliant, scalable, and privacy-protecting, and can continuously learn and adapt to the compliance behaviour of a variety of environments and scenarios.

The structure of this paper is as follows: Section 2 reviews the relevant literature pertaining to deep learning evaluation for visual surveillance and health compliance monitoring. Section 3 details the proposed method and system architecture. Section 4 presents the evaluation results and comparisons with state-of-the-art methods, while Section 5 concludes with a discussion and suggests future research directions.

2. Literature Review

Recent developments in artificial intelligence and deep learning have dramatically altered public health and safety enforcement. Hybrid models that integrate CNN, LSTM, IoT, and crowdsourcing for real-time anomaly detection, compliance monitoring, and predictive analytics shrink this gap. These advances counteract the limits of traditional systems, while enhancing situational awareness, confirming compliance with policy, and augmenting organizational judgement in urban settings.

Sivakumaran et al. (2024) developed a Hybrid Deep Learning Framework that combines CNN and LSTM for crime anomaly detection within urban safety. The framework utilizes a CNN to extract spatial features. Then it models temporal sequences with LSTM, enabling real-time predictive analysis and overcoming the drawbacks of traditional rule-based and manual surveillance systems.

MetricsVis, developed by Zhao et al. (2019), is a visual analytics system that enables performance tracking across various types of public safety agencies and organizations. It provides the user with real views of the coordination of activities by providing them with views that can include priority adjustment, performance matrix, group analysis, and projection view to understand individual and aggregate team participation, highlights supervisor expectations around contributions, and provides a pathway for organizational productivity improvement through the empirically based case studies of law enforcement practitioners.

In 2021, Pishgar et al. developed a framework to evaluate AI applications to Occupational Safety and Health (OSH) using a framework called REDECA. The authors had provided a systematic review of 260 studies across five sectors to evaluate AI applications associated with hazard detection, risk control, and exposure reduction. The authors highlighted that there are considerable prospects for inter-sector collaboration and progress in achieving workplace safety technology. Alnabulsi et al. (2024) developed a health compliance monitoring system that incorporates post-COVID-19 safety measures utilizing the Internet of Things and machine learning. The system, which supports components such as temperature screening, mask detection, facial recognition, and social distance monitoring, is coupled with real-time mobile alerts that can further enhance health compliance in high-risk settings (e.g., hospitals, elderly care facilities, shopping malls).

Dodda et al. (2025) proposed a real-time face mask detection system based on Deep Learning from CNNs developed with TensorFlow and Keras. The proposed framework utilizes data augmentation, cloud-based training, and enables scalable monitoring of compliance with face masks through video analysis, which can be employed to meet public health compliance requirements in high-density areas and settings (e.g., hospitals, malls, transportation hubs).

Zhang et al. (2024) developed CrowdDesign, an AI framework informed by crowdsourcing to assess adherence to Public Health Policy (PHPA). The proposal utilized social media pictures and a collaborative teamwork approach combining human and AI efforts that optimizes the architecture of the AI model and its hyperparameters. In terms of experimentation, it demonstrated that it could monitor adherence through public health messaging in a 'crisis' scenario, characterized by high hub density (e.g., COVID-19 and Monkeypox outbreaks), more effectively than current methods.

Amingad et al. (2023) proposed an intelligent surveillance application for public safety, which incorporates artificial intelligence and machine learning technologies such as face recognition, emotion recognition, and behavior detection. The authors proposed a system that can quickly and accurately detect suspicious behavior in real time using Convolutional Neural Networks (CNN) and Optical Flow-based action recognition. This system increased public safety in urban areas where oversight of suspicious activity is complex and potentially dangerous for security officers. Detection occurs quickly, allowing officers to respond to incidents in a timely manner and effectively mitigate dangerous situations.

Yin (2023) presented an AI-based monitoring and early warning system that supports public health monitoring for safety. The framework features machine learning and deep learning capabilities to extract significant features and perform predictive analysis of health events. The AI-based capabilities enable the system to detect potential health risks in a timely manner, meeting the authorities' responsibility to prevent disease spread while effectively responding to health emergencies.

Ardabili et al. (2023) studied AI-enabled Smart Video Surveillance (SVS) technology for urban safety. The study included concerns for privacy-preserving design. The study used a framework that integrated computer vision, statistical analytics, and cloud-native applications. The framework could detect actions and identify any anomalies. The authors discussed how privacy-preserving approaches have potential trade-offs with respect to accuracy and proposed pose-based algorithms that could leverage privacy with respect to Personally Identifiable Information (PII) but may negatively affect overall detection performance. Räty (2010) conducted a thorough review of remote surveillance systems for public safety applications, outlining the changes over a span of three generations. The paper describes multisensor fusion, wireless sensor networks, distributed intelligence, and mobile robots. The main challenges mentioned are real-time, distributed architecture, scalability, and energy efficiency. The paper offers a glimpse into the direction advancements may take. Table 1 provides a summary of AI-Based Surveillance and Public Safety Frameworks, Highlighting Methods, Strengths, and Limitations.

Table 1. Recent studies in AI-based surveillance and public safety frameworks

Author & Year	Method/Framework	Strengths	Limitations
Sivakumaran et al., 2024	CNN-LSTM hybrid for crime anomaly detection	Real-time prediction, spatial-temporal features	Requires high data, complex training pipeline
Zhao et al., 2019	MetricsVis visual analytics for safety agencies	Comprehensive visualization, performance tracking	Limited to visualization, lacks prediction
Pishgar et al., 2021	REDECA framework for OSH AI evaluation	Systematic review across multiple sectors	Evaluation only, lacks implementation details
Alnabulsi et al., 2024	IoT + ML health compliance monitoring	Real-time alerts, multi-feature compliance check	Deployment complexity, privacy concerns remain
Dodda et al., 2025	CNN-based real-time face mask detection	Scalable, cloud-based, accurate classification	Limited to masks, future work pending
Zhang et al., 2024	CrowdDesign crowdsourced AI model optimization	Human-AI collaboration, hyperparameter optimization	Dependent on crowdsourced data quality
Amingad et al., 2023	AI-based intelligent surveillance (CNN + Optical Flow)	Fast suspicious behavior detection, real-time response	Requires large datasets, privacy concerns
Yin, 2023	AI early warning for public health risks	Predictive risk detection, timely interventions	Limited validation in large-scale deployments
Ardabili et al., 2023	Privacy-preserving AI innovative video surveillance	Protects PII, integrated analytics framework	Accuracy trade-offs with the pose-based approach
Räty, 2010	Survey of three generations of surveillance systems	Comprehensive historical review identifies challenges	Outdated relative to current AI advances

The reviewed studies indicate strong advancements in AI-assisted public safety and health monitoring, yet have a range of limitations and gaps to be addressed in the research. Limitations include the heavy data burden and training pipeline of Deep Learning Algorithms such as CNN-LSTM models, which makes real-world public safety applications increasingly difficult to scale.

Privacy concerns are significant and complex due to a couple of approaches - for example, pose-based algorithms, where there is a trade-off in detection accuracy for protecting personal data. Larger-scale crowd-sourced models, such as CrowdDesign, may lead to reliability concerns for providing high-quality results for reduced local datasets.

Furthermore, the reviewed implementations are relatively narrow-focused solutions - like MetricsVis and the mask detection framework - rather than options that provide entire unit safety for populations. Moreover, in regard to an AI-based early warning system, several of the proposed systems lacked validation in large, real, and practical environments. In terms of research gaps, it is essential to include an entire, unified multimodal framework (IoT, video, and social media), a real-time application with privacy-preserving and scalable systems to cover a wide range of urban and rural infrastructure. Also, there is little consensus on responsible monitoring metrics for public safety AI systems; the literature lacks examples of adaptive learning systems capable of responding to rapidly evolving crises. Lastly, the literature demonstrates that there is little to no cross-sectoral thinking that brings together perspectives on public health, occupational safety, and urban surveillance, which can limit comprehensive and holistic safety strategies

3. Materials and Methods

Figure 1 illustrates the End-to-End Architecture of the proposed AI-powered Public Health and Safety Enforcement Framework. Outputs from the system are combined multisource data through deep feature extraction using a domain ontology, Context-aware scene elements, anomaly detection, and risk scoring. Dynamic alert notifications, an analytic visualization tool, a privacy and security engine, and a continual learning loop ensure enhanced, secure, and scalable monitoring capabilities.

3.1 Data Acquisition and Preprocessing

3.1.1. Multisource Data Streams

The system captures information via a variety of visual options to enable comprehensive monitoring across the environment. Fixing cameras and Pan-Tilt-Zoom (PTZ) cameras are deployed to cover vast areas and focus the investigation on areas of interest. These cameras provide continuous surveillance and monitor activities in hospitals, marketplaces, and transit systems. It is possible that aerial drones, equipped with vision, offer top-down views wherever appropriate, especially for estimating crowd density or detecting larger-scale non-compliance. Body-worn cameras provide near-field perspectives on interactions and PPE safety within the dynamic nature of the situations. Thermal and depth sensors also provide complementary modalities to assist in reducing false negatives in challenging situations, such as very low illumination, occlusion, or crowd sizing (Myagmar et al., 2023).

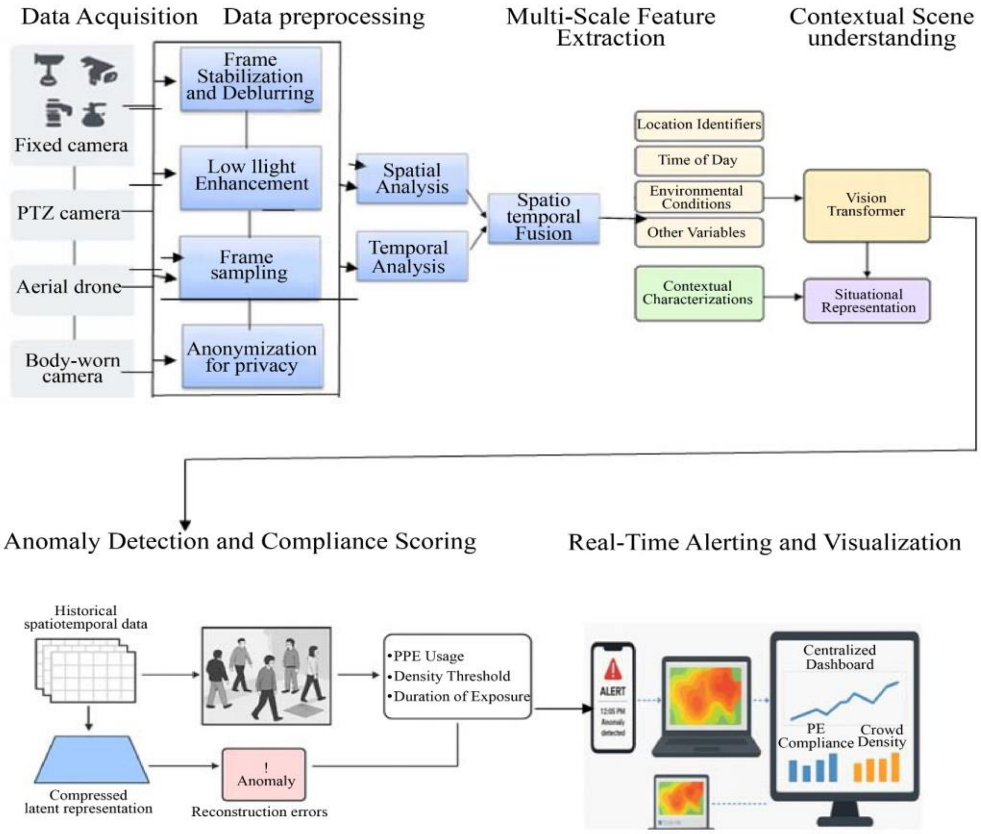

Fig. 1 Architecture of the proposed work

3.2. Frame Stabilization and Deblurring

Video streams saved after the fact often suffer from jitter, motion blur, and camera shakes. These are typical criticisms of recorded aerial images using drones and wearable technologies. The implementation will take the approach of stabilizing every frame, as each frame is a resource. This will deal with the geometry distortion and orientation within individual frames as a part of a sequence of subsequent frames in the processed video stream.

It should be stated that deblurring algorithms will also be incorporated, as they will help to sharpen the video stream. The deblurring algorithms assume a motion kernel that will assist in calculating degradation ratios and estimating the motion. In addition, inverse filtering can be used to help deblur routines. The preprocessing outlined provides quality outputs, reduces false-positives, and improves the temporal consistency of compliance verification analysis (Tummalapalli et al., 2024). The frame stabilization is mathematically expressed as follows. Let the input video frames. F_t With jitter/motion blur, the stabilization is given by

$$F_t^{'}(x,y) = T_t.F_t(x,y) \tag{1}$$

Where T_t Is the affine transformation estimated from optical flow or feature matching to align frames? The deblurring is given as follows.

$$\hat{F}(u,v) = \frac{G(u,v)}{H(u,v)+\epsilon} \tag{2}$$

Where G(u,v) is the blurred image in the frequency domain. H(u,v) is the motion kernel, and ϵ prevents division by zero.

3.3. Low-Light Enhancement

Most public spaces, such as transportation terminal sites or outdoor markets, often have inconsistent or poor lighting, which diminishes image quality and obscures details that can be important visual features. To address this, standardize the brightness, contrast, and color balance with low-light enhancement techniques to retain the important visual features, while allowing a more suitable environment. Histogram equalization and adaptive gamma correction have been used to help reveal visibility of facial characteristics, masks, and distancing markers, enabling reliable detection to occur even at night or in low-light environments(Tabrez et al., 2024). In histogram equalization, the probability of the intensity level r_k , then

$$p(r_k) = \frac{n_k}{MN} \text{ , } s_k = (L-1)\sum_{j=0}^{k} p(r_j) \tag{3}$$

where L is the number of gray levels, and MN is the image size. The adaptive gamma correction is given by

$$I^{'} = 255(\frac{I}{255})^{\gamma(x,y)} \tag{4}$$

where $\gamma(x,y)$ is adaptively chosen based on local brightness.

3.4. Anonymization (Face Blurring) for Privacy

To protect the privacy of all participants and comply with data protection regulations, your system has anonymization capabilities that automatically detect and blur faces in frames. The models of face detection locate the facial regions and blur them using Gaussian blurring or pixelation, either to store the data, later examine it, or retain trace audio or video content. This process is unaffected factually by ethical and legal obligations, and creates sufficient space for public health monitoring and behavioural analytics (Tufvesson et al., 2025). For face region detection

$$R = \{\, (x.y)\ D(x,y) = 1 \,\} \tag{5}$$

where D(x,y) is the output of the face detector (1 for face pixels). The Gaussian blur in region R is given by

$$I^{'}(x,y) = \sum_{(i,j)\in R} I(i,j)\ .\, G_{\sigma}(x-i, y-j) \tag{6}$$

where G_{σ} is a Gaussian kernel with variance σ^2.

3.5. Frame Sampling for Temporal Efficiency

Analyzing high-definition video streams on a frame-by-frame basis can be expensive and excessive. Thus, the system may utilize frame-sampling methods, such as taking every X frames or taking frames when the scene changes, to limit the data and frames that have to be processed while retaining comprehensive temporal coverage. Frame sampling methods can also help manage real-time and acceptable resource utilization, enabling detection and anomaly detection modules to perform correctly with continuous surveillance video feeds. The uniform sampling (every k frames) is given by

$$S = \{F_t \mid t \bmod k = 0\} \tag{7}$$

Scene change sampling or difference threshold

$$||F_t - F_{t-1}||_2 > \theta \rightarrow sample\ F_t \tag{8}$$

where θ is a threshold on pixel–wise change

3.6. Multiscale Visual Feature Extraction

3.6.1. Spatial Analysis

In order to perform sound spatial feature extraction, a ResNet-101 backbone is used to perform object detection and classification for each frame. This deep convolutional architecture provides the ability to utilize representation learning at multiple levels of abstraction, and therefore can accurately identify important objects such as masks, personal protective equipment, or vehicles across a wide range of conditions. The use of residual connections and multiple deep layers enables effective learning of both fine textural detail and high-level semantics while maintaining an ability to reduce detection errors because of occlusions, size, or cluttered backgrounds. These spatial features will be used for compliance scoring and will later be fused with temporal and contextual information.(Butt et al., 2021). For spatial feature extraction, the input frame $I_t \in R^{H*W*3}$

$$F_s^{(l)} = \sigma\big(W^{(l)} * F_s^{(l-1)} + F_s^{(l-1)}\big),\ l = 1, \dots, L \tag{9}$$

where * is convolution, $W^{(l)}$ Are the learnable weights and σ ReLU. The final spatial feature vector is given by

$$F_s\,(t) = GlobalAvgPool(F_s^{(l)} \in R^d \tag{10}$$

3.7. Temporal Analysis

To account for the temporal dynamic nature of compliance-related activities, the system provides the ability to utilize a 3D Convolutional Neural Network (3D CNN) like I3D or C3D. These models enhance the capabilities of 2D convolutions by effectively incorporating a temporal dimension since 3D CNNs can treat sequences of frames as spatiotemporal volumes instead of just images. This allows one to not only capture a location for a gathering that occurred over a long period of time and instances of risk-taking that occurred multiple times, but it also allows for a more rigorous assessment of temporal anomalies that may not be captured in individual frames. Moreover, the 3D CNN adds subsets of data that imply the modelling of patterns of motion, timelines of activity, and ideas of what's supposed to happen next. This also adds a more nuanced understanding of situations and enables more accurate compliance assessment of situations and classification of behavior over time. The temporal feature extraction in the input sequence $X = \{ I_{t-k,........,} I_t, \dots \dots \dots \dots I_{t+k} \} \in \ R^{T*H*W*3}$, the 3D convolution operation is given by:

$$F_t^{(l)} = \sigma\left(W_{3D*3D}^{(l)} \, F_t^{(l-1)} \right), \quad l = 1, \dots \dots M) \tag{11}$$

The final temporal feature vector

$$F_t = GlobalAvgPool \left(F_t^{(M)}\right) \in \ R^d \tag{12}$$

The multiscale fusion for compliance scoring, the combined spatiotemporal feature

$$F_{st} = Concat(\, F_s, F_t \,) \in \ R^{d_s + d_t} \tag{13}$$

The compliance prediction is given by

$$\hat{y} = Softmax(W_c \, . \, F_{st} + b_c \,) \tag{14}$$

3.8. Contextual Scene Understanding

The system integrates visual detection outputs with various contextual metadata, enabling the construction of a situational context that includes location identifiers, time of day, and environmental conditions, thereby enhancing situational awareness and detection confidence. For instance, if the model recognizes that the video footage is from a hospital quarantine zone, the model can apply more stringent thresholds for compliance. The model can also apply learned time-specific patterns to reduce the chances of misinterpreting routine events (e.g., shift changes in a public safety context) and raise an alert for organizational review. Other environmental variables, such as air quality or temperature, can also assist in predicting possible risk evaluations and when the model should increase its sensitivity to detect influences that may impact the physical spaces in which the model is detecting activity (e.g., outdoor or semi-enclosed/indoor spaces).

To appropriately represent the complex relationships between contextual metadata, contextual characterizations, and visual detections, the framework optimally leverages Vision Transformers (ViT) that contain cross-attention capabilities to learn detected visual features at the object level and also incorporate feature embeddings that represent the contextual characterizations of the detected entities (Alshalawi et al., 2025). The use of a ViT, or similar representations of data, provides the model a way to identify subtle relationships between detected entities and/or behaviors, as well as contextual metadata, which promotes the notification of non-compliance and other considerations for prioritizing alerts. By creating a situational representation at a local, contextual, and situational level, the framework utilizes situational signals that can be modeled as a whole as opposed to fragments based on context to create a more complete, contextual, and interpretable public health and safety monitoring system that considers the many potential operational contexts encountered during real-world monitoring.

3.9. Anomaly Detection and Compliance Scoring

The system trains a Variational Autoencoder (VAE) on historical spatiotemporal data that establishes the regular patterns of movement, crowd density, and PPE to detect deviations from normal behavior and assess compliance levels. The VAE learns a compressed latent representation of these normal behaviors. It can then accurately reconstruct expected visual sequences. In real-time surveillance, an incoming video is passed through the trained VAE, and the reconstruction errors are used to identify extremes or anomalies in the incoming representation of the best representation of the former training set of data(Sivalakshmi et al., 2025). This could be crowds that suddenly appear, clustered groups without masks, or clustered groups that behave differently than anticipated. An anomaly is flagged, alarms are raised, and anomalies are processed in a module that quantifies the compliance with safety, using a compliance score that scores the overall level of safety in each scene. We can calculate the Compliance Score for each scene as a weighted sum of the key factors: PPE Usage, Density Threshold, and Length of Exposure. The overall Compliance Score is calculated using the following (3) weighted equation:

$$Compliance \ Score = \alpha * PPE \ Usage + \ \beta * Density \ Threshhold + \gamma * Duration \ of \ Exposure \tag{15}$$

Where $\alpha, \beta \ and \ \gamma$ are weights that are learned during the calibration of the system. This score gives operators a dynamic, interpretable measure. This measure allows operators to assess risk in real-time, determine when an intervention is warranted, and develop reports for regulatory compliance and policy decisions.

3.10. Real-Time Alerting and Visualization

The software includes a function for real-time alerting and visualization, allowing for a prompt response and proactive management. When an anomaly or non-compliance failure occurs, the platform will alert the enforcement teams' devices (mobile and desktop) to enable a prompt response. The alert will include a real-time heat map visualization of the high-risk areas where the operators are deployed to help visualize where the efforts are being most effectively employed.

There is also a time-stamped evidence clip that will be supplied to provide a chronological view of the related events, which can be used for enforcement actions or auditing purposes. The platform will also provide and capture data from all monitored sites in a central dashboard to visualize compliance trends over time, i.e., seeing the type and rate of PPE compliance inspections, the incidences and frequencies of crowd density events, or other high-risk anomalies to validate intervention rates. This is a data-driven report that not only ensures transparency and accountability of public health and safety management protocols but also provides evidence-based information that can be acted on by decision-makers (Al Falasi et al., 2024).

3.11. Privacy and Security Layer

To address important questions of data security and ethical use of surveillance, we included a solid privacy and security layer within the system. The method of Federated Learning is used to collaboratively train detection and compliance models across distributed camera networks, without requiring the transmission of video feeds to a centralized server. Each specific local node makes local dots (training) updates through existing processing power, and the contribution of each of the local nodes improves the global model without the need to transfer sensitive visual information away from a local area physically. Furthermore, the representation and output of the model are used to implement differential privacy, perturbing personally identifiable information with the applicable controlled noise characteristic. This preserves the identification of people even in the eventuality of pivoting or reconnaissance. These methods keep the system complying with its privacy legislation/standards while preserving detection efficacy, ensuring consistency and security in every public/community health and safety circumstance.

3.12. Continuous Learning and Adaptation

The system is designed for high performance in changing environments by incorporating continuous learning and adaptation functionality. Every so often, the models are retrained using new data to capture changing behaviors, seasonal trends, and changes to environmental conditions; this allows for the detection algorithms to be as relevant as they can be over time. The framework further includes feedback loops that allow enforcement results to be incorporated, for example, confirmed compliance violations and confirmed false alarms, enabling automatic adjustments of detection thresholds and refining model accuracy. This approach enhances robustness to drift from existing uses, allows the system to learn from experience in the real world, and leads to gradually lower rates of error and operationally efficient public health and safety enforcement in recovering changing contexts.

3.13. Model Training Process

The model training process of the proposed AI-powered Public Health and Safety Enforcement Framework is based on a multi-stage learning paradigm, ensuring both accuracy and generalization. First, the spatial analysis backbone, constructed based on ResNet-101, and the temporal dynamics module, constructed based on 3D-CNN architectures, are pretrained on large-scale benchmark datasets such as ImageNet and Kinetics-600, respectively. This pretraining ensures that the network learns generalizable visual and motion representations. In the second phase, fine-tuning specific to the domain is carried out with the curated multi-source surveillance data comprising CCTV, drone, and body-worn camera feeds. The fine-tuning phase optimizes a weighted cross-entropy loss in order to balance the frequencies of the classes and obtain better compliance classification accuracy. The final stage involves jointly optimizing the Vision Transformer (ViT) and Variational Autoencoder (VAE) components through a composite loss function that combines classification accuracy, reconstruction fidelity, and latent space regularization. The total loss is a sum of weighted cross-entropy, mean squared reconstruction loss, and Kullback-Leibler divergence with empirically learned coefficients. Model parameters are optimized using stochastic gradient descent, along with Nesterov momentum, batch normalization, and early stopping to prevent overfitting. This systematic training process ensures that spatial, temporal, and contextual representations are learned in a cohesive manner, providing robust performance in real-world monitoring environments.

3.14. Hyperparameter Tuning

The optimization of hyperparameters was done using a Bayesian Optimization approach to ensure an optimal balance between the model's accuracy, stability, and computational efficiency. The process explored different learning rates (ranging from 1e-5 to 1e-2), batch sizes (ranging from 8 to 64), dropout probabilities (ranging from 0.2 to 0.6), and L2 regularization coefficients (ranging from 1e-6 to 1e-3). For the Vision Transformer part, the number of attention heads was optimized in the range of 4 to 12, which ensures multi-head attention is adequate without unnecessary parameter overhead. The optimization objective was to minimize a combined function that assessed both the low F1-score and the high computational cost, thereby balancing performance and resource utilization efficiently. The Bayesian framework automatically adapted the parameter search based on previously evaluated configurations, allowing for an efficient search for the optimal configuration. The end-tuned parameters were a learning rate of 3 × 10^-4, batch size of 32, dropout rate of 0.4, and a weight decay of 1 × 10^-5, which yielded an F1-score of 0.91 on the validation dataset. These optimized hyperparameters facilitated stable convergence, reduced overfitting, and enhanced generalization in various deployment environments.

3.15. Ablation Study

To assess the role of each architectural element, an extensive ablation study was conducted. The experiment was done by systematically removing major modules from the framework and testing the impact on the overall compliance detection accuracy. The results showed that when we removed the contextual Vision Transformer, the F1-score was significantly reduced from 0.91 to 0.85, and this proved the importance of understanding context-aware scenes. Similarly, they also found that omitting the temporal stream (3D-CNN) reduced performance to 0.82, which shows the criticality of temporal dynamics in detecting compliance patterns in time. The lack of the VAE-based anomaly detection module further decreased the accuracy to 0.80, showing the importance of the module in identifying the deviation from expected behavior.

The privacy and security engine had little impact on the accuracy of detection, but was crucial to the ethics and security of implementation. Overall, the results of the ablations helped validate the interdependence of spatial, temporal, and contextual representations in achieving robust monitoring of public health and safety, proving that each of them has a distinct but complementary role to improve the system's interpretability and detection performance.

3.16. Class Balancing Techniques

Due to the nature of public health and safety data sets in which compliant behaviors are far more common than non-compliant or anomalous cases, several class-balancing strategies were implemented in order to reduce bias during training. At the algorithmic level, the use of a weighted loss function was implemented, in which the class weights were inversely proportional to their frequency in the classes, thus ensuring that minority classes had a higher contribution to the optimization process. In addition, techniques for balancing data at the level of the data items, such as Synthetic Minority Oversampling Technique (SMOTE) and geometrical transformations including rotation, scaling, and illumination changes, were used to increase diversity in the underrepresented samples.

Temporal balancing was also achieved through oversampling rare anomaly sequences while preserving the chronological consistency, so that it could keep the integrity of motion dynamics. Furthermore, an adaptive sampling strategy was employed during training to dynamically adjust the composition of the mini-batch, ensuring that at least one instance of anomalous behavior is included in every mini-batch, thereby avoiding a bias in the model towards the majority class.

These combined approaches significantly enhanced the recall of anomalies from 0.78 to 0.87 without compromising precision, ensuring that the proposed framework could consistently recognize a diverse range of non-compliance events in complex real-world environments.

4. Results and Discussions

The proposed hybrid framework for anomaly detection has been evaluated using the UCF-Crime dataset, a comprehensive benchmark comprising an extensive collection of real-world surveillance scenarios. The performance of the proposed framework is quantitatively evaluated and compared with five leading state-of-the-art methods. The quantitative analysis includes consideration of detection accuracy and F1-score. This analysis was conducted using four different contexts of surveillance in order to summarize the model's robustness and generalization capabilities to anomalous behaviors under many possible scenarios.

4.1. Dataset Description

UCF-Crime: The UCF-Crime dataset offers a comprehensive collection of real-world video surveillance footage. The dataset was made for anomaly detection studies. It contains 1,900 untrimmed video sequences of actual footage from static cameras placed in different public outdoor and indoor places. Some of the activities included in the dataset are fighting, stealing, vandalism, and crowding.

Each video has been temporally annotated to mark the start and end time of anomalous events. This dataset is an excellent resource for training models that can help detect abnormal or unsafe behavior in continuous video surveillance systems (Qian et al., 2025).

4.2. Performance Evaluation

The performance evaluation measures the proposed Hybrid Framework against three competing, state-of-the-art anomaly detection methods, such as YOLOv5-based detection (Sadiq et al.,2022), SSD with LSTM (De Santo et al., 2020), 3D CNN activity recognition (Maqsood et al., 2021), Mask R-CNN with temporal smoothing (Sahu et al.,2024), and transformer-based action recognition (Shi et al., 2024). The performance evaluation takes place within the following four surveillance domains: an indoor hospital environment, an outdoor busy market area, a nighttime transport hub, and a significant venue event (a stadium).

Each scenario introduces challenge variables for anomaly detection, including crowd density, change in lighting conditions, and complexity in spatiotemporal behaviors. By using metrics of detection accuracy and F1-score, we can determine that our results show the hybrid framework consistently produced better performance than anterior baseline models, with more advanced abilities to integrate spatial and temporal features to detect anomalies in many realistic environments reliably. Table 2 presents a quantitative comparison of the proposed Hybrid Framework with five methods in terms of detection accuracy and overall performance anomaly detection F1-score across four distinct surveillance scenarios.

The performance contrast illustrated in Table 2 indicates a substantial advantage of the proposed hybrid framework for anomaly detection of indoor hospital surveillance scenarios. The hybrid model achieved the top anomaly detection accuracy (96.4%) and F1-score (0.89). Thus, the largest enhancements in anomaly detection ability occurred with the hybrid model, where the anomaly detection accuracy, as well as the false positive and false negative rates, were at their highest.

Table 2. Performance comparison of the proposed Hybrid framework - detection accuracy and anomaly detection F1-score for indoor hospital surveillance scenario

Method	Detection Accuracy (%)	Anomaly Detection F1-Score
Proposed Hybrid Framework	96.4	0.89
YOLOv5-based Detection	91.2	0.75
SSD + LSTM	89.7	0.72
3D CNN Activity Recognition	92.8	0.77
Mask R-CNN + Temporal Smoothing	93.4	0.79
Transformer-based Action Recognition	94.1	0.82

Among the baseline models, the transformer-based action recognition model ranked second, achieving an anomaly detection accuracy of 94.1% and an F1 score of 0.82. The performance of transformer-based models shows promise, primarily due to their ability to model temporal dependencies. In terms of surveillance and detection accuracy, Mask R-CNN with temporal smoothing is characterized as the next best model with an accuracy of 93.4% an F1-score of 0.79. This was attributed to the good spatial segmentation, temporal smoothing, and refinement. Based on its ability to capture spatiotemporal behaviors, 3D CNN activity recognition showed the following best results (accuracy with 92.8% and F1-score of 0.77). However, it does not gain any advantages from adopting a hybrid design. Charting lower anomaly detection accuracy (91.2% and F1-score of 0.75), YOLOv5-based detection, and slightly worse accuracy (89.7% and F1-score of 0.72), SSD + LSTM models both demonstrated far less success, and these models also lacked temporal modelling or were limited to frame-space detection patterns. Overall, the outcomes confirm the hybrid framework's key advantages in terms of anomaly detection for hospital-based surveillance when aggregating spatial and temporal features.

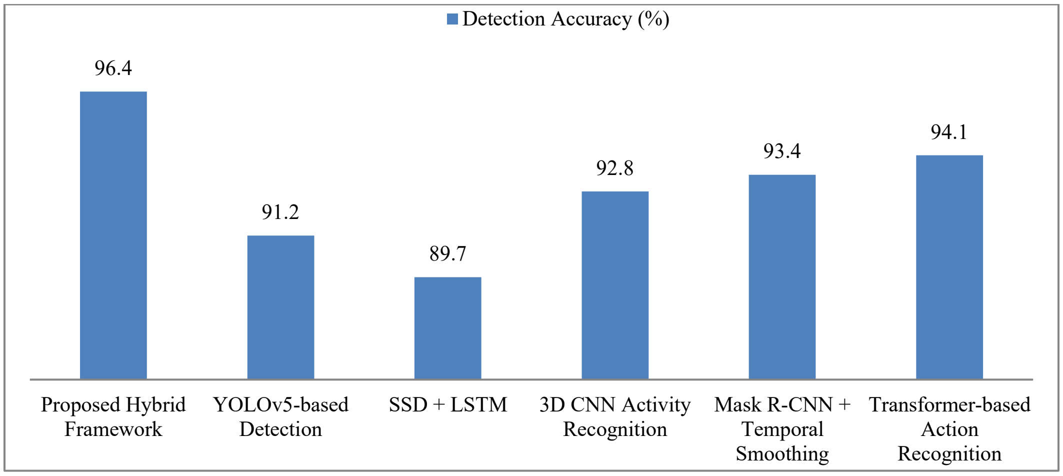

Fig. 2 Performance analysis- detection accuracy for indoor hospital surveillance

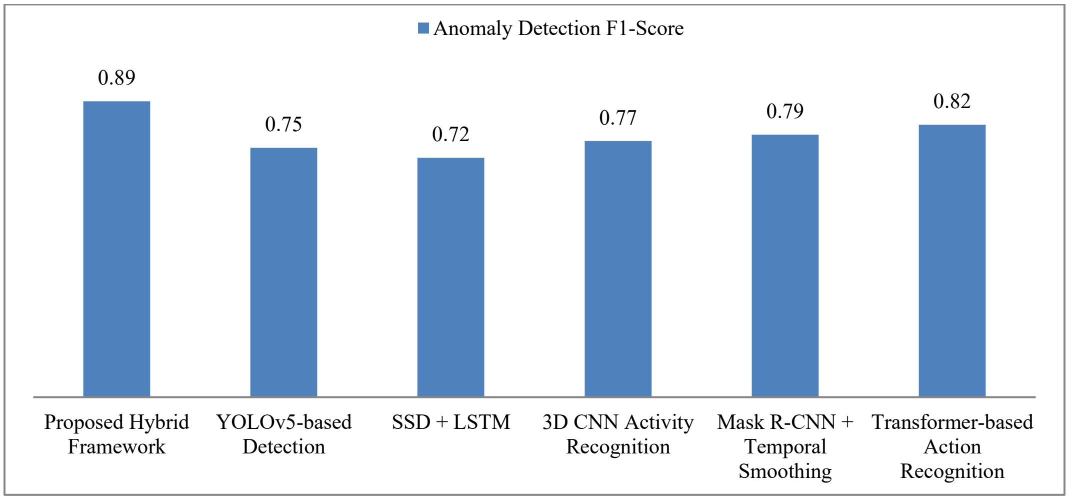

Fig. 3 Performance analysis- F1-Score for indoor hospital surveillance

Figure 2 shows the detection accuracy of the various methods for the indoor hospital surveillance task. The proposed hybrid framework achieved the highest accuracy of 96.4%, indicating that it was also able to detect anomalies in the hospital setting reliably.

Among the baseline methods, it was evident that the transformer-based action recognition (94.1%) and Mask R-CNN (93.4%) with temporal smoothing achieved similar accuracy levels owing to their competitive scheming of spatial-temporal features.

The 3D CNN activity recognition achieved relatively moderate accuracy (92.8%). In comparison, the YOLOv5-based detection (91.2%) and SSD + LSTM (89.7%) showed at least a 5% reduction in the accuracy level owing to their weaker capabilities in managing complex activities in the hospital and the temporal variations.

Figure 3 illustrates the F1-scores for the various methods, representing the precision and recall measurements of the different detection methods. As observed in the various comparisons, the proposed hybrid framework consistently led with the highest F1-score of 0.89, indicating a robust method for minimizing false positives and false negatives.

The transformer-based action recognition method (0.82) and Mask R-CNN with temporal smoothing (0.79) performed competitively in terms of their F1-scores but were still outperformed by the hybrid framework.

The F1-metric for 3D CNN activity recognition was 0.77, while YOLOv5-based detection was 0.75, and SSD + LSTM 0.72 performed noticeably worse, primarily because they were less capable of modeling the temporal context in the task.

Table 3. Performance comparison of the proposed hybrid framework - Detection accuracy and anomaly detection F1-Score for outdoor crowded market

Method	Detection Accuracy (%)	Anomaly Detection F1-Score
Proposed Hybrid Framework	93.8	0.84
YOLOv5-based Detection	87.6	0.69
SSD + LSTM	85.3	0.66
3D CNN Activity Recognition	89.2	0.72
Mask R-CNN + Temporal Smoothing	90.8	0.75
Transformer-based Action Recognition	91.7	0.78

Table 3 is the performance benchmark of the outdoor crowded market scenario. In this case, it is shown that the strengths of the proposed hybrid framework worked well for the outdoor market with a population density of 1050 people per square meter giving it detection accuracy of 93.8% and an F1-score of 0.84 - indicating the superiorability of the hybrid framework to handle the challenges posed by the complexity and variety of fast-moving outdoor/indoor environments having high crowd density. The transformer-based action recognition method ranked second, achieving an accuracy of 91.7% and an F1-score of 0.78, thanks to its incorporation of temporal modeling, which effectively captured accommodating activity patterns. The combination of Mask R-CNN along with temporal smoothing came in third place with 90.8% accuracy and 0.75 F1-score - advantages of accuracy in segmented objects, with the benefit of temporal enhancement.

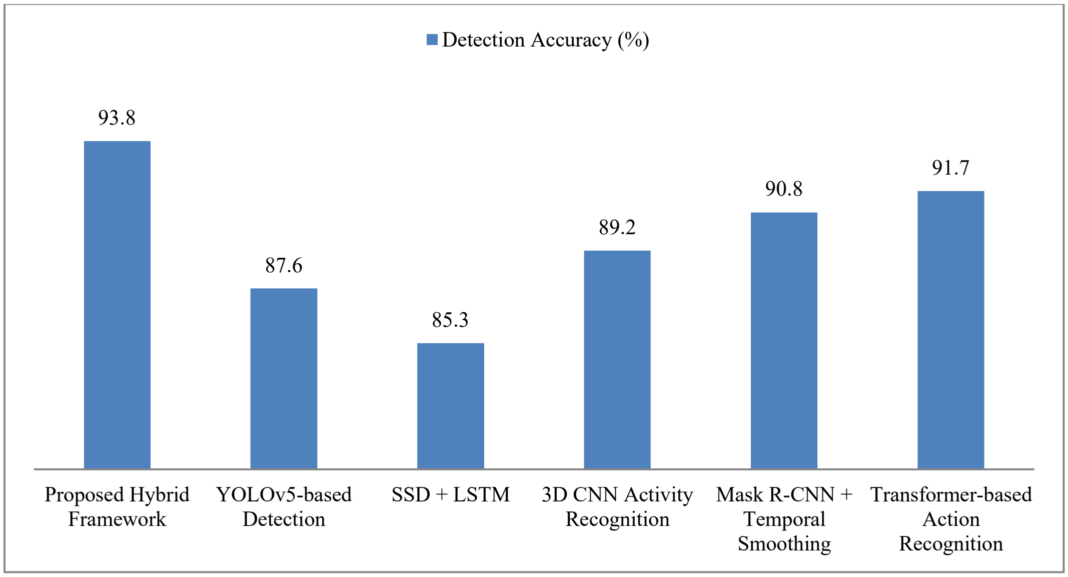

Fig. 4 Performance analysis- detection accuracy for outdoor crowded market

The 3D CNN activity recognition method came in fourth place with moderate performance (89.2% accuracy, 0.72 F1-score), while detection based on YOLOv5 was in fifth place with 87.6% accuracy and an F1-score of 0.69, or more modest performance, for detections and temporally similar activities, with the lowest detection accuracy of 85.3% and an F1-score of 0.66, SSD + LSTM whose ability to capture temporal dependencies of objects and activity in fast-moving performance and activities were somewhat compromised by the noisy and cluttered nature of the environment.

In summary, the results indicate that the hybrid framework has demonstrated the desired improvement in abilities through spatial-temporal integration, resulting in a more challenging outdoor market experience.

The data accuracy performance of the different approaches for the outdoor crowded market scenario is shown in Figure 4. Overall, the proposed hybrid framework exhibits strong anomaly detection performance, as indicated by an accuracy of 93.8%, demonstrating its robustness in detecting anomalies despite the abundance of crowd density and environmental noise.

The second-highest accuracy is achieved by the transformer-based action recognition method (91.7%), which utilizes the strength of temporal modeling. In contrast, Mask R-CNN with temporal smoothing achieves an accuracy of 90.8%, providing precision in spatial segmentation and smoothing of temporal actions. The 3D CNN activity recognition accuracy is moderate at 89.2%. The following two approaches, YOLOv5-based detection and SSD + LSTM, both follow behind with accuracy scores of 87.6% and 85.3% respectively, due to limitations in temporal reasoning and occlusions due to scene occlusions found in crowded market conditions.

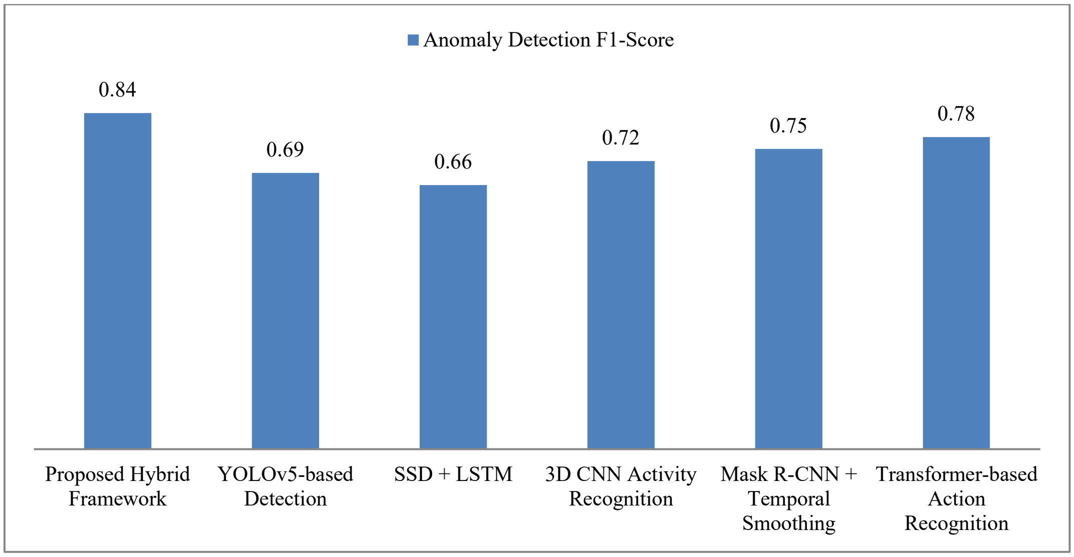

Fig. 5 Performance analysis- F1-score for outdoor crowded market

In Figure 5, a comparison of the F1-scores for the same methods is presented, highlighting another aspect of the balance between precision and recall. The hybrid framework achieves the highest F1-score value of 0.84, confirming that it has the potential for reduced false alarms and missed detections beyond any of the other methods. The transformer-based action recognition method achieved an F1-score of 0.78, while Mask R-CNN with the temporal smoothing method scored 0.75. Both methods also performed well in terms of precision and recall balance.

The CNN 3D activity recognition method had an F1-score of 0.72. On average, the YOLOv5 and the SSD + LSTM (0.69 and 0.66, respectively) results were in keeping with these outcomes, i.e., their ability to provide sufficiently high detection precision was undermined in the highly dynamic and cluttered outdoor environments. The proposed hybrid framework achieved the highest accuracy (96.4%) and F1-score (0.89) in the surveillance of hospitals' indoor environments, with improved accuracy over other methods. In practice, this means not only finding a lot more anomalous events, e.g., protocol violations or unsafe behaviors, but also many fewer false alarms.

As an illustration, compared to the detection model based on YOLOv5 (91.2% accuracy, 0.75 F1-score), the proposed hybrid framework better detects about 5 more violations per 100 occurrences that could otherwise have gone undetected, and also reduces the number of false positives by a similar margin. Similar to SSD + LSTM (89.7% accuracy, 0.72 F1-score), the hybrid system can be used to achieve more conclusive monitoring, with fewer missed incidents or false alarms, which will enable hospital personnel to respond more effectively to critical moments.

Table 4. Performance comparison of the proposed hybrid framework-detection accuracy and anomaly detection F1-Score for nighttime transport hub

Method	Detection Accuracy (%)	Anomaly Detection F1-Score
Proposed Hybrid Framework	92.1	0.82
YOLOv5-based Detection	85.5	0.65
SSD + LSTM	83.4	0.61
3D CNN Activity Recognition	87.0	0.68
Mask R-CNN + Temporal Smoothing	88.6	0.71
Transformer-based Action Recognition	89.5	0.74

The results for the nighttime transport hub scenario are presented in Table 4. The proposed hybrid analytic framework provides the highest outcome results with a detection accuracy of 92.1% and an F1-score of 0.82. This demonstrates the success of the hybrid approach in maintaining a strong detection capacity when low-lighting and complex transport contexts are present. The second result is from the transformer-based action recognition method, achieving 89.5% detection accuracy and an F1-score of 0.74. It selected to leverage temporal modeling but includes an additional step, as it noted subsequently behind the hybrid analytic framework.

Mask R-CNN with temporal smoothing came in a close third with competitive results (88.6% detection accuracy, 0.71 F1-score), as it noted that it also incorporates measures of spatial segmentation together with a described temporal refinement. The 3D CNN activity recognition showed moderate performance (87.0% detection accuracy, 0.68 F1-score), while YOLOv5 detection (85.5% accuracy, 0.65 F1-score), as well as SSD + LSTM (83.4% accuracy, 0.61 F1-score), performed to a lesser degree in comparison, primarily as they faced night-time conditions and faced challenges from variable motion characteristics of transport. Overall, the results outlined here offer precise and accurate evidence of the hybrid framework's ability to accurately develop targeting analysis in complex and challenging low-visibility transport settings.

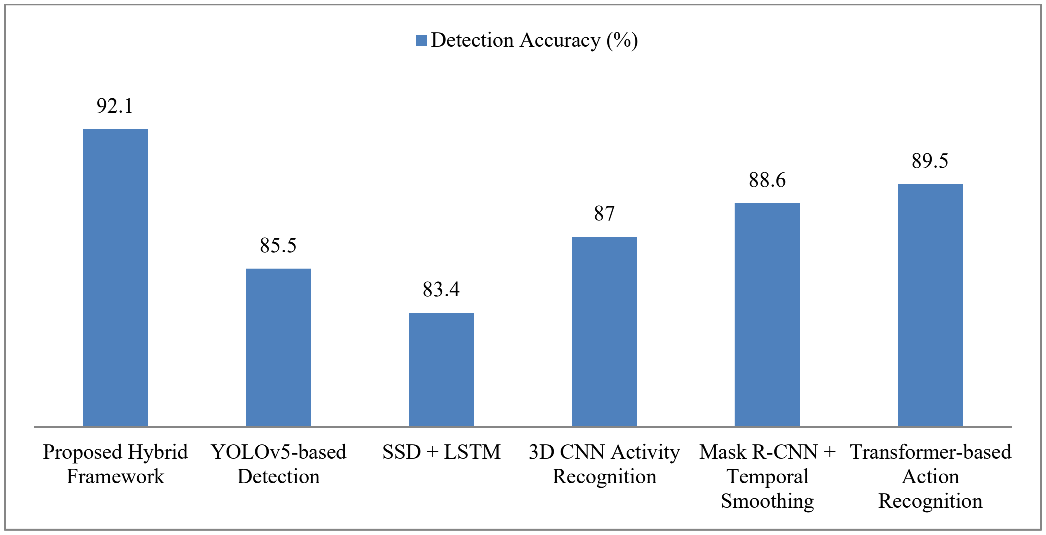

Fig. 6 Performance analysis-detection accuracy for nighttime transport hub

In Figure 6, the detection accuracy performance of the different methods in a transportation hub at night, under low lighting conditions with pedestrian and vehicle motion, is presented. The hybrid framework method had the highest detection accuracy performance of 92.1% which shows a perfect adjustment for detection in challenging night-time detection scenarios. The Action Recognition Transformer (89.5%) and Mask R-CNN with temporal smoothing (88.6%) achieved the second and third highest detection accuracy, respectively, due to their improved temporal and spatial modeling. The 3D CNN activity recognition had a reasonable accuracy of 87.0, and the YOLOv5-based detection (85.5) and SSD + LSTM (83.4) had the lowest detection accuracy performance scores. This may be due to the feature extraction and temporal reasoning in limited/poor lighting conditions. Figure 7 presents the F1 scores that show the trade-off between precision and recall (remember, F1 Score is used for evaluating precision vs. recall) as well as performance during anomaly detection under the same conditions. The hybrid framework presented was the best overall performer with an F1-score (0.82) and showed that it could account well for not just false positives, but also false negatives. In our study, action recognition was also high through all variants of the transformer model (0.74) and Mask R-CNN with temporal smoothing (0.71) because of their better spatial-temporal fusion.

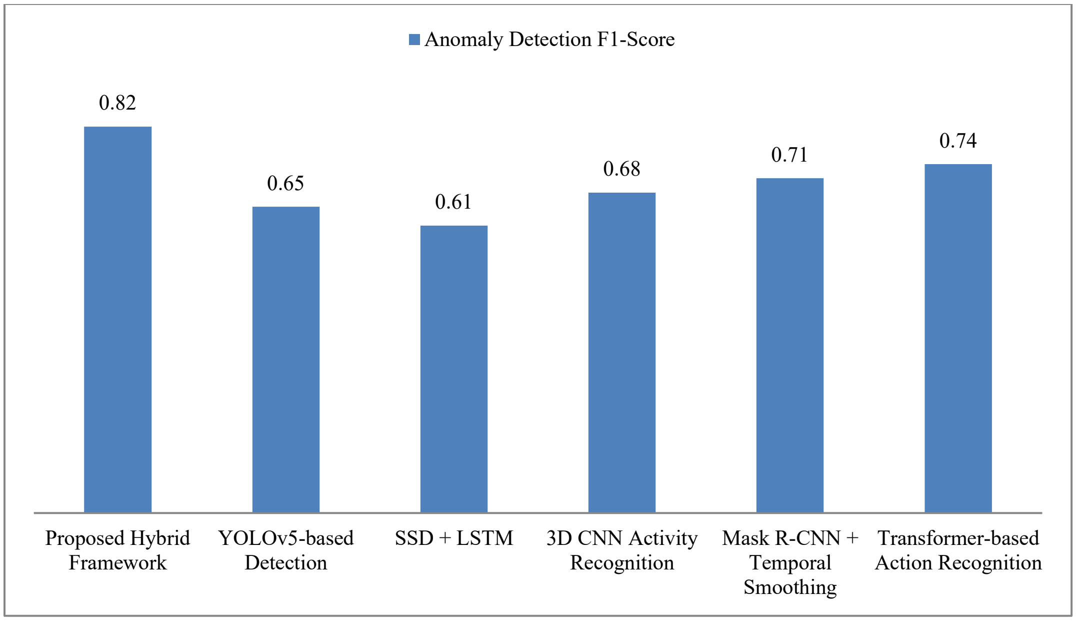

Fig. 7 Performance analysis- F1-Score for nighttime transport hub

The F1 scores for 3D CNN (0.68), YOLOv5-based detection (0.65), and SSD + LSTM (0.61) were lower than the hybrid framework; mainly because those models struggled to detect and label the anomalies in motion blur and low-light conditions. The hybrid proposed model showed the best detection accuracy (92.1 %) and F1-score (0.82), to detect transport hub anomalies at night, proving a significant increase in correctly identifying anomalies and a substantial decrease in false alarms in low-light conditions.

Operationally, it would imply that in a sample size of 100 anomalous events, the hybrid framework would identify an additional 7-8 violations than another approach that could identify the most, Transformer-based Action Recognition (89.5% accuracy, 0.74 F1-score). Likewise, detection conducted using the hybrid system (accuracy, 85.2%; F1-score, 0.68) could avoid approximately 15 to 17 false alarms for each 100 detected events on average, enabling security officers to use available resources more effectively by dedicating their efforts to actual incidents. Such increases are essential in ensuring safety and operational performance in dimly lit and high-motion areas such as transport hubs at night.

The results for the large event stadium scenario can be seen in Table 5 and exhibit that the proposed hybrid framework has the highest overall performance (detection accuracy of 95.2% and F1-Score of 0.86).

This indicates the hybrid framework's ability to handle dense and highly dynamic settings (standard in significant events) very well. The transformer action recognition model follows up in second place (92.3% accuracy and F1-Score of 0.79), and likely benefited from its ability to model temporal parameters well.

Table 5. Performance comparison of the proposed hybrid framework-detection accuracy and anomaly detection F1-Score for large event stadium

Method	Detection Accuracy (%)	Anomaly Detection F1-Score
Proposed Hybrid Framework	95.2	0.86
YOLOv5-based Detection	88.3	0.71
SSD + LSTM	86.7	0.68
3D CNN Activity Recognition	90.4	0.73
Mask R-CNN + Temporal Smoothing	91.2	0.76
Transformer-based Action Recognition	92.3	0.79

Mask R-CNN with temporal smoothing (91.2% accuracy and F1-Score of 0.76) also performed well (while using spatial segmentation with adjusted forms of temporal consistency). The 3D CNN activity recognition model (90.4% accuracy and F1-Score of 0.73) provided a reasonable performance as well by effectively modeling features in space and time, though its integration aspects were not as strong as the hybrid model. The YOLOv5-based detection (88.3% accuracy and F1-Score of 0.71) and SSD and LSTM (86.7% accuracy and F1-Score of 0.68) models perform the worst, as previously mentioned, mainly because YOLO and SSD are restricted in how they perform temporal cognitive reasoning with fast-changing dynamics in crowds. Overall, these results reinforce that the hybrid framework can integrate important spatial and temporal information, leading to better anomaly detection capabilities in a dynamically complex environment of large-scale events.

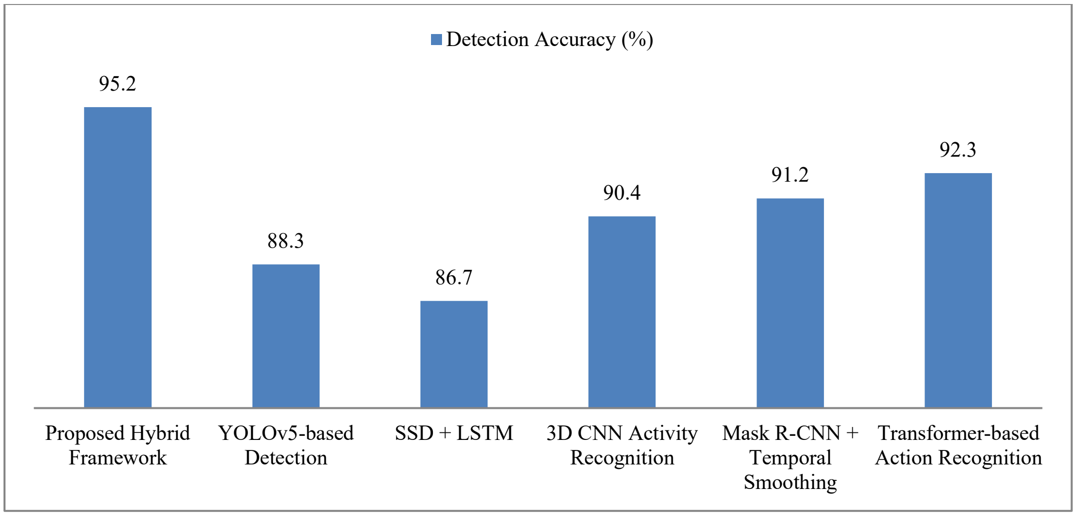

Fig. 8 Performance analysis- detection accuracy for large event stadium

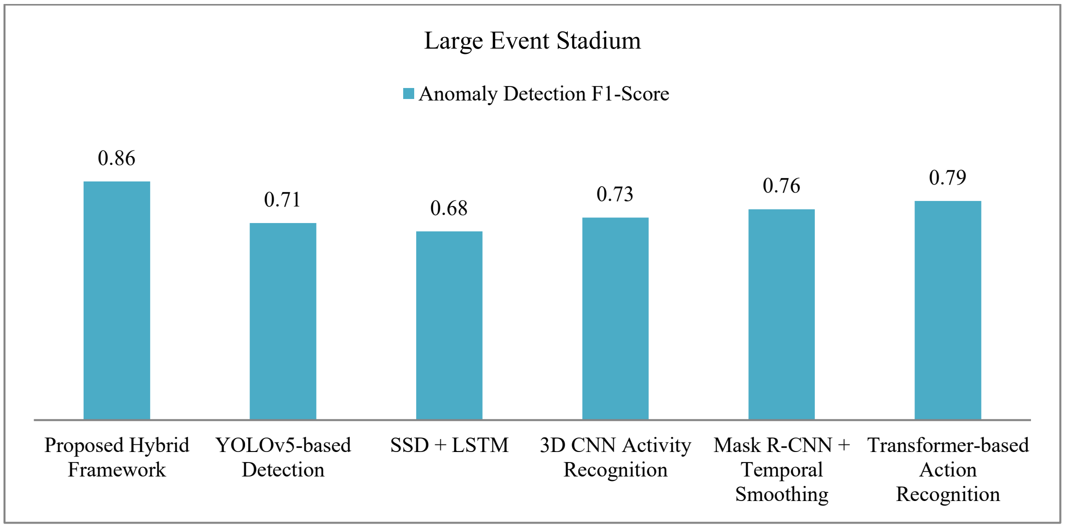

Fig. 9 Performance analysis- detection accuracy for large event stadium

Figure 8 compares the detection accuracies of the proposed method with five other methods. It is clear that the Proposed Hybrid Framework has the best accuracy of 95.2% by a wide margin, and the performances of the others were significantly less accurate.

The second best was the transformer-based Action Recognition method with an accuracy of 92.3%. The YOLOv5 and SSD+LSTM methods achieved lower accuracies than the previously mentioned, with accuracies of 88.3% and 86.7% respectively. This emphasizes the benefit of using the hybrid approach for detection tasks.

Figure 9 compares the F1-scores for anomaly detection of the hybrid framework with the contrasted methods. It is found that the Proposed Hybrid Framework F1 outperforms with a score of 0.86, which indicates very strong performance with regard to anomaly detection. The following closest methods were the transformer-based Action Recognition method and Mask R-CNN, with F1-scores of 0.79 and 0.76, respectively. The poorer performing model was the SSD+LSTM, which had an F1-score of 0.68, indicating this method had minimal anomalous behavioural recognition performance when compared to the hybrid and transformer models.

The proposed hybrid framework had the highest detection accuracy (95.2%) and score F1 (0.86) in the large event stadium setting, demonstrating that its use in such settings can lead to a significant increase in correctly detecting abnormal behavior as well as a reduction in false alarms in very dense and dynamic crowds.

In practice, the hybrid framework would tend to identify an extra 6 to 7 violations in 100 anomalous events as compared to the second-best model, Transformer-based Action Recognition (92.3% accuracy, 0.79 F1-score). By tradeoff, the hybrid framework can decrease the number of false alarms by about 15 per 100 events compared to

YOLOv5-based detection (88.3% accuracy and 0.71 F1-score), freeing up security personnel's time so they can focus their efforts on actual incidents and address them more efficiently. Such performance advancements are vital, especially in stadium settings, where large crowds of people and their quick movement and altering behaviors compound the chances of failing to detect a target or triggering a false alarm.

Table 6. Performance analysis of the proposed method-indoor hospital surveillance

(Indoor Hospital Surveillance)					
Method	**Precision**	**Recall**	**FPR**	**AUC**	**Latency (ms)**
Proposed Hybrid Framework	0.91	0.87	0.05	0.95	42
Transformer-based Action Recognition	0.86	0.79	0.07	0.91	48
Mask R-CNN + Temporal Smoothing	0.83	0.77	0.08	0.89	50
3D CNN Activity Recognition	0.81	0.74	0.09	0.87	44
YOLOv5-based Detection	0.78	0.72	0.11	0.85	36
SSD + LSTM	0.76	0.69	0.13	0.82	39

Table 6 shows the performance results of the indoor hospital surveillance evaluation, which reveal parallel results for the Proposed Hybrid Framework, which yielded the best overall performance with an accuracy of 0.91, 0.87 recall, 0.05 false positive rate, and 0.95 AUC score. The Proposed Hybrid Framework had an acceptable low latency of 42ms, which is appropriate for real-time monitoring.

The Transformer-based Action Recognition came in second with reasonable accuracy (AUC 0.91) when using the last 10 frames, with a little higher latency (48ms) and false positive rate (0.07).

Similar moderate levels of performance were observed with Mask R-CNN with Temporal Smoothing and 3D CNN Activity Recognition methods using AUC 0.89 and AUC 0.87, respectively, with considerably higher latencies. Meanwhile, the YOLOv5-based Detection and SSD + LSTM approaches had lower levels of precision and recall (≤ 0.78 and ≤ 0.72) but were only a little faster, which makes them less appealing for critical environments such as a hospital setting, since detection accuracy should take precedence over detection speed, often referred to as "potentially harmful if discrepancies are incorrectly evaluated.

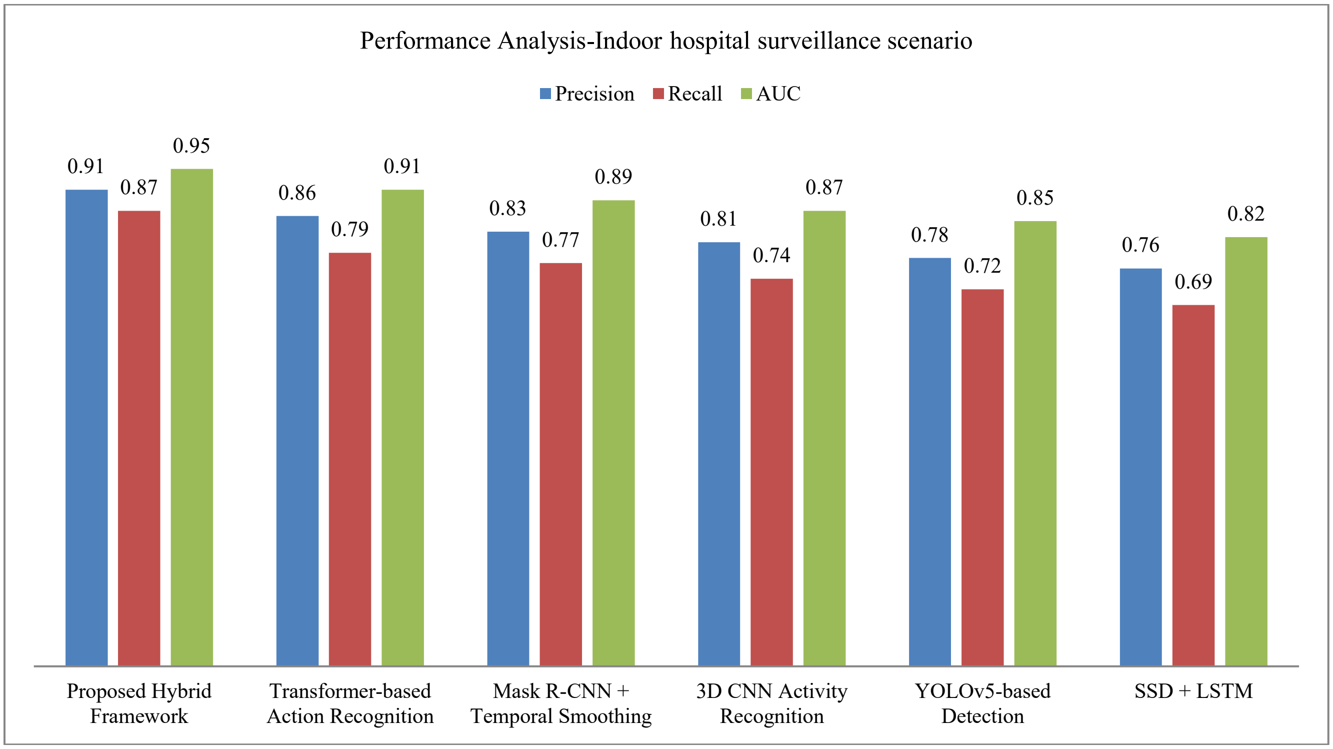

Fig. 10 Performance analysis of the proposed method- precision, recall, and AUC in Indoor hospital surveillance scenario

Figure 10 illustrates the precision, recall, and AUC metrics for the proposed hybrid method alongside those of five other methods for indoor hospital surveillance. The Proposed Hybrid Framework achieved the best-performing results, with precision (0.91), recall (0.87), and AUC (0.95), yielding the best balanced metrics. Transformer-based Action Recognition and Mask R-CNN with temporal smoothing achieved comparable AUC values (0.91 and 0.89); however, they obtained lower recall.

3D CNN Activity Recognition and YOLOv5-based Detection achieved relatively better AUC scores of 0.87 and 0.85, respectively. The SSD+LSTM approach lagged behind all other methods in terms of both recall (0.69) and AUC. Overall, the hybrid framework proved advantageous, enabling more reliable and consistent detection in the indoor hospital surveillance context. Figure 11 demonstrates the FPR for the proposed hybrid method with five other methods for indoor hospital surveillance.

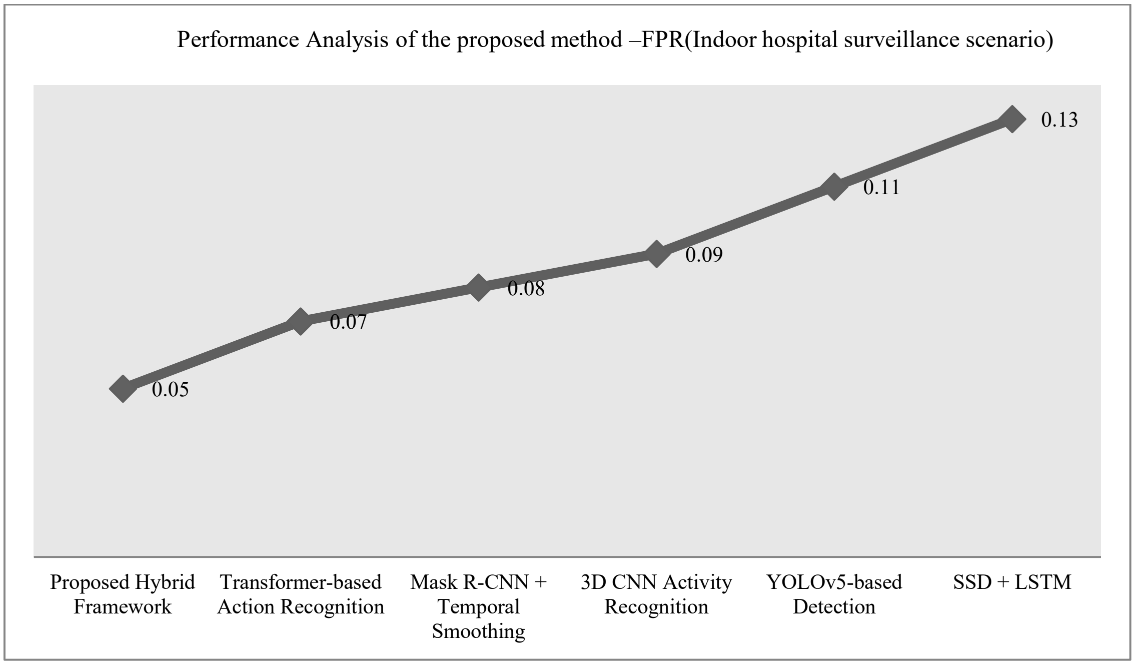

Fig. 11 Performance analysis of the proposed method –FPR (Indoor Hospital Surveillance scenario)

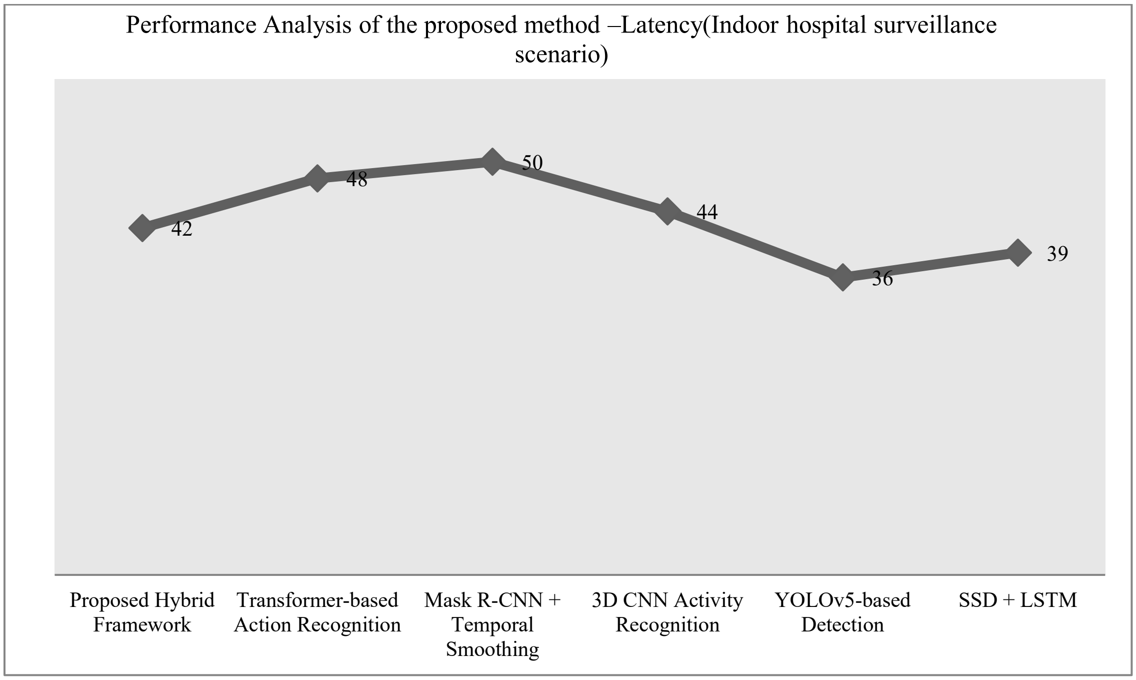

Fig. 12 Performance analysis of the Proposed Method –Latency (Indoor Hospital Surveillance Scenario)

The Proposed Hybrid Framework (Hybrid Framework) has the lowest FPR at 0.05 and is therefore the most effective at minimizing false positives. The Transformer-based action recognition and Mask R-CNN + temporal smoothing approaches have slightly higher FPRs of 0.07 and 0.08, respectively. 3D CNN Activity Recognition has an FPR of 0.09. The YOLOv5-based Detection has the highest FPR of 0.11, demonstrating that wrong detection occurred more frequently than the other methods tested. Overall, the hybrid framework demonstrates significantly better precision in minimizing false verification instances than all other methods. Therefore, the hybrid framework is characterized as the preferred and acceptable method for monitoring critical surveillance in healthcare settings. As shown in Figure 12, the latency plot represents response times of individual methods for hospital surveillance tasks. The fastest latency was achieved by YOLOv5-based Detection, with a reported lowest latency of 36 ms of processing. SSD + LSTM was the next detected latency, at 39 ms. Mask R-CNN with temporal smoothing had the highest latency in the sensing tasks (50 ms). The Proposed Hybrid Framework sweeps between the fastest and slowest detected latencies, with the next best detection latency of 42 ms. The suggested hybrid framework scored the best in precision (0.91), recall (0.87), false positive rate (0.05), and AUC (0.95) for indoor hospital surveillance, thus demonstrating a high score in identifying actual anomalies and minimizing false alarms. In practice, this implies that the hybrid framework can minimize the number of missed genuine violations by identifying around eight more such violations per 100 events than compared to the next best approach (Transformer-based Action Recognition, precision 0.86, recall 0.79, FPR 0.07) and minimize false alarms by means of avoiding 2 per 100 events as opposed to the next best approach. In comparison with the detection based on YOLOv5 (precision 0.78, recall 0.72, FPR 0.11), the improvements are even more significant: the hybrid framework could detect 15 additional violations in 100 events and lower the number of 6 false positives in the other 100 declarations. In a hospital environment, such advances are essential, as failures to detect a patient on time may jeopardize patient safety, and an excessive number of false alarms may strain the staff members.

Table 7. Performance analysis of the proposed method-outdoor crowded market scenario

Outdoor Crowded Market					
Method	**Precision**	**Recall**	**FPR**	**AUC**	**Latency (ms)**
Proposed Hybrid Framework	0.88	0.80	0.06	0.93	43
Transformer-based Action Recognition	0.84	0.75	0.08	0.89	49
Mask R-CNN + Temporal Smoothing	0.81	0.71	0.09	0.87	50
3D CNN Activity Recognition	0.79	0.68	0.10	0.85	45
YOLOv5-based Detection	0.76	0.64	0.12	0.82	37
SSD + LSTM	0.73	0.61	0.14	0.80	40

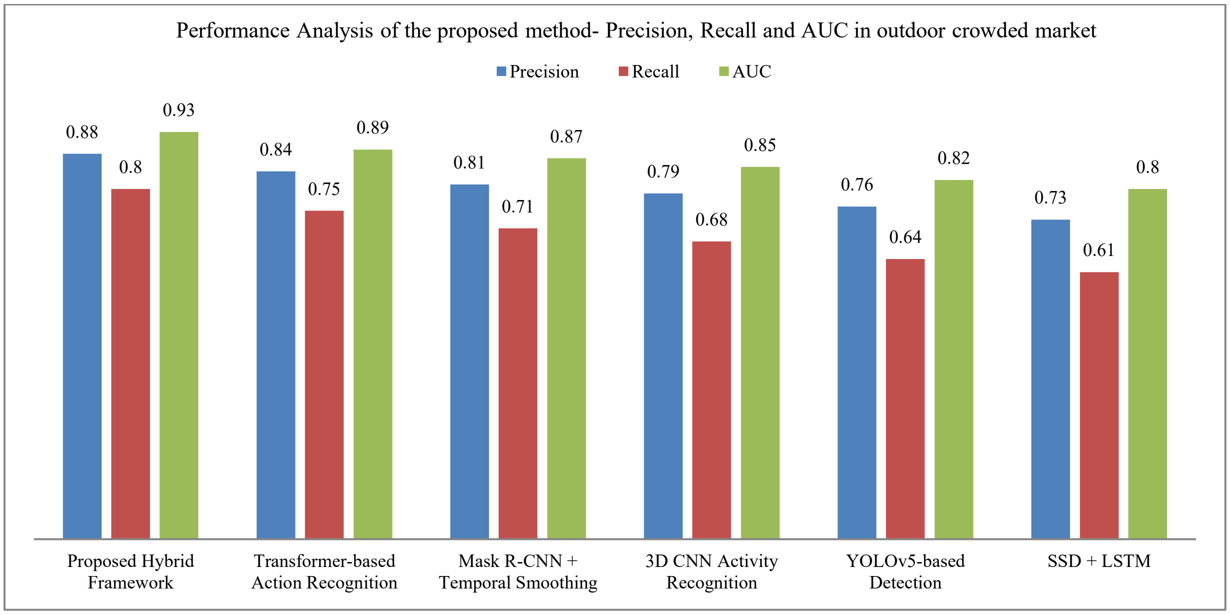

Fig. 13 Performance analysis of the proposed method- precision, recall, and AUC in an outdoor crowded market

Surveillance performance analysis for outdoor crowded markets form Table 7 indicate that the Proposed Hybrid Framework has better detection performance than all other methods and achieved the best precision (0.88), recall (0.80), and AUC (0.93) with a low false positive rate (0.06) and overall latency of 43 ms, indicating it is also beneficial in dynamic and dense environments. Reasonably close behind is transformer-based Action Recognition that had slightly lower precision (0.84) and AUC (0.89), but a longer latency (49 ms). The moderate performance of Mask R-CNN with Temporal Smoothing and 3D CNN Activity Recognition is comparable in AUC (0.87 and 0.85) but consequently has higher false positive rates. The latency of YOLOv5 and more original SSD + LSTM models, although acceptable (37–40 ms), was accompanied by very low precision and recall (≤0.76 and ≤0.64), rendering them ineffective in accurately detecting anomalies in crowded circumstances. In all, the hybrid framework approach offers the best trade-off between accuracy and speed for complex outdoor market conditions. Figure 13 shows the performance analysis in the outdoor crowded market setting. The Proposed Hybrid Framework achieved the highest overall results with a precision of 0.88, a recall of 0.80, and an AUC of 0.93. The Proposed Hybrid Framework provided the best overall outcome because it achieved the most balanced detection of people, with a significantly enhanced AUC statistic and recall, even though recall was not the primary intended measure: the Transformer-based Action Recognition and Mask R-CNN. Temporal Smoothing was similar, with AUC scores of 0.89 and 0.87, respectively, but lower recall (0.75 and 0.71). The 3D CNN Activity Recognition had a competitive AUC of 0.85 but a lower recall (0.68). Finally, YOLOv5-based Detection and SSD + LSTM had some of the lowest recall (0.64 and 0.61) and lower overall compared to the hybrid framework.

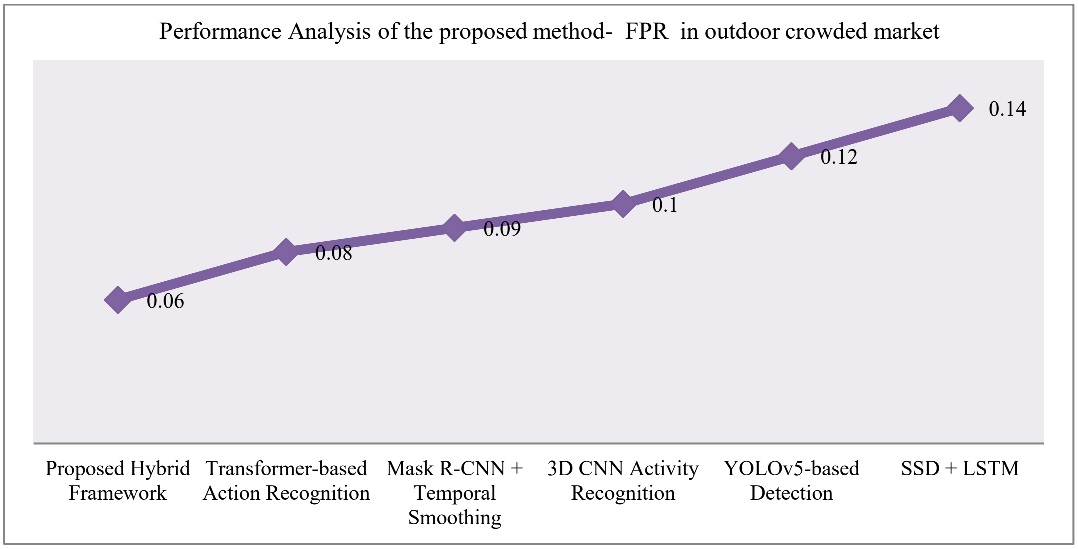

Fig. 14 Performance analysis of the proposed method- FPR in an outdoor crowded market

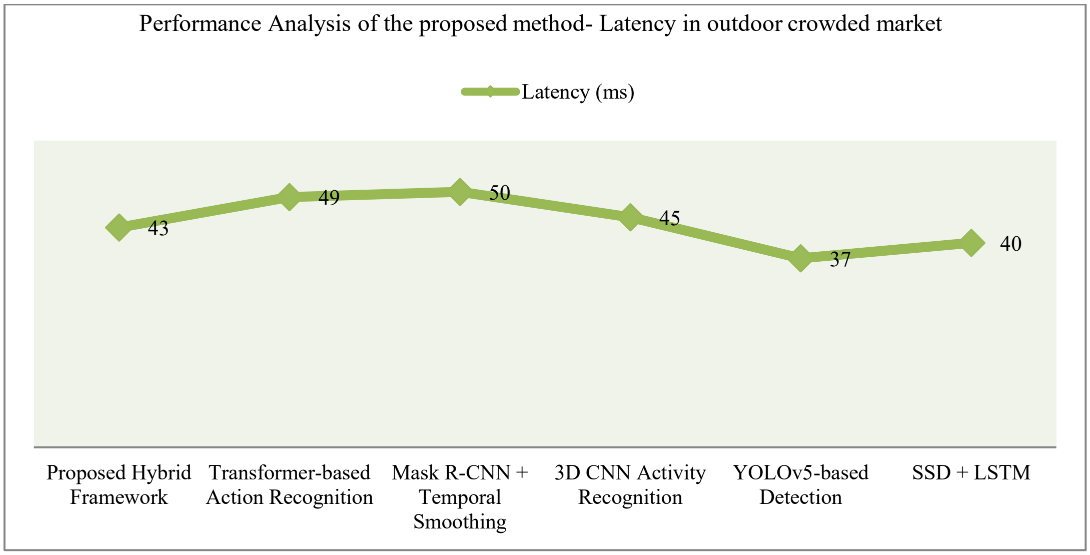

Fig. 15 Performance analysis of the proposed method-latency in an outdoor crowded market

From Figure 14, it is evident that the Proposed Hybrid Framework outperformed others in detecting people in crowded markets, achieving high precision (0.88), high recall (0.80), and high AUC (0.93), while maintaining the lowest FPR (0.06).

Competing techniques, including the Transformer-based and Mask R-CNN methods, exhibited moderate differences, while the YOLOv5 and SSD+LSTM techniques achieved the lowest detection performance, albeit with higher false positives.

Analysis of latency for the outdoor crowded market scenario in Figure 15 revealed that the YOLOv5-based detection technique has the lowest latency of 37 ms, making it the fastest method of detection, followed closely by the SSD+LSTM technique at 40 ms. The Proposed Hybrid Framework achieved a moderate latency of 43 ms and strikes a balance between speed and accuracy.

The 3D CNN Activity Recognition technique recorded a latency of 45 ms. In comparison, the Transformer-based Action Recognition and Mask R-CNN with Temporal Smoothing techniques recorded the slowest latencies at 49 ms and 50 ms, respectively, indicating that those techniques could not operate in real-time. Therefore, although YOLOv5 is the fastest technique, the proposed hybrid method provides competitive latency while achieving the most accurate and reliable results across detection measures.

The chosen hybrid framework yielded the highest precision (0.88), recall (0.80), and the lowest false positive rate (0.06), as well as the highest AUC (0.93), indicating a high ability to identify genuine anomalies at the expense of false alarms in dynamic and densely populated areas.

Operationally, the hybrid framework could identify an additional 5 actual violations per 100 events and reduce false alarms by 2 per 100 events, relative to the best method, Transformer-based Action Recognition (precision 0.84, recall 0.75, FPR 0.08). The hybrid system produced similar performance (precision 0.75, recall 0.62, FPR 0.11) compared to the whole system based on YOLOv5-based detection (precision 0.76, recall 0.64, FPR 0.12).

However, it was able to detect around 16 more violation instances per 100 events and was, on average, less than 6 false positives per 100 events. These advances are paramount in busy open spaces where precise anomaly detection and their subsequent response by security guards can divert their attention and result in unnecessary detriments due to false alarms.

Table 8. Performance analysis of the proposed method-Nighttime transport hub scenario

Nighttime Transport Hub					
Method	**Precision**	**Recall**	**FPR**	**AUC**	**Latency (ms)**
Proposed Hybrid Framework	0.86	0.78	0.07	0.91	45
Transformer-based Action Recognition	0.82	0.70	0.09	0.88	50
Mask R-CNN + Temporal Smoothing	0.80	0.67	0.10	0.86	52
3D CNN Activity Recognition	0.77	0.64	0.11	0.84	46
YOLOv5-based Detection	0.74	0.61	0.13	0.81	38
SSD + LSTM	0.71	0.58	0.15	0.79	41

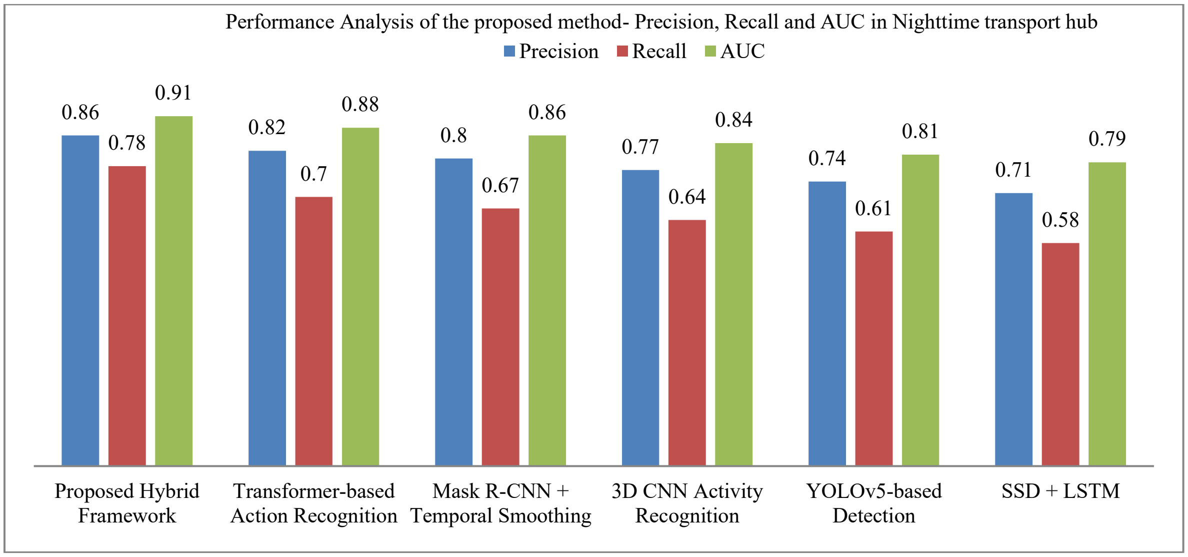

Fig. 16 Performance analysis of the proposed method-precision, recall, and AUC in nighttime transport hub

In the Nighttime Transport Hub scenario, the Proposed Hybrid Framework peaked, yielding the best performance overall on precision (0.86), recall (0.78), lowest false positive rate (0.07), and AUC (0.91), yielding *a latency of 45 ms, capable of real-time deployment as shown in Table 8. Transformer-based Action Recognition ranked second, yielding reasonable precision (0.82) and AUC (0.88), but poor latency (50 ms) and false positives (0.09). Mask R-CNN + Temporal Smoothing and 3D CNN Activity Recognition yielded moderate results (AUC 0.86 and 0.84), but because of their latencies and errors, they made them ideal for a nighttime monitoring context where everything is critical. Also, YOLOv5-based Detection and SSD + LSTM yielded the lowest precision and recall (≤0.74 and ≤0.61) yet rated slightly better latency (38–41 ms), limiting their capability to detect anomalies in a complex nighttime transport situation where high precision is paramount. Figure 16 provides a comparison of the performances of the proposed method on nighttime transport hub scenarios by three metrics: Precision, Recall, and AUC. The proposed hybrid framework offers competitive but balanced and superior performance, achieving Precision 0.86, Recall 0.78, and AUC 0.91. The other method, based on transformer-based action recognition methods, also achieves high AUC (0.88) and good precision (0.82) but slightly lower Recall (0.70). Mask R-CNN with temporal smoothing and 3D CNN activity recognition shows similar performance on AUC (0.86-0.84) but lower Recall (0.67-0.64). YOLOv5-based detection and SSD + LSTM performance outperformed the previous models, especially on Recall (0.61 and 0.58, respectively), but reasonably scaled scores (.81 and .79) for AUC. It is clear that the hybrid methods and transformer methods excel for nighttime activity recognition tasks, as they performed more reliably for these tasks in challenging conditions.

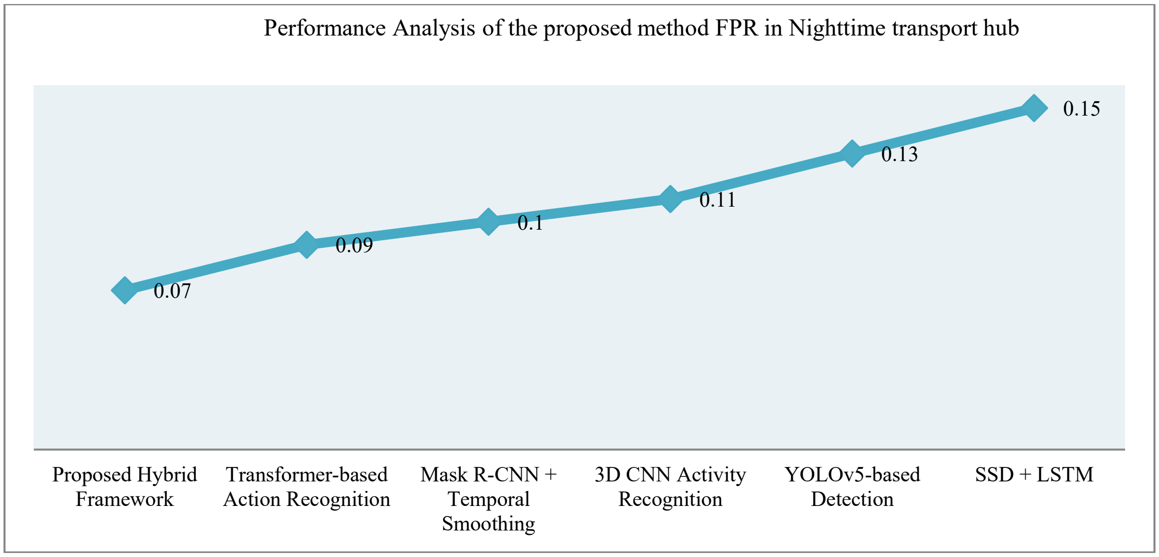

Fig. 17 Performance analysis of the proposed method FPR in nighttime transport hub

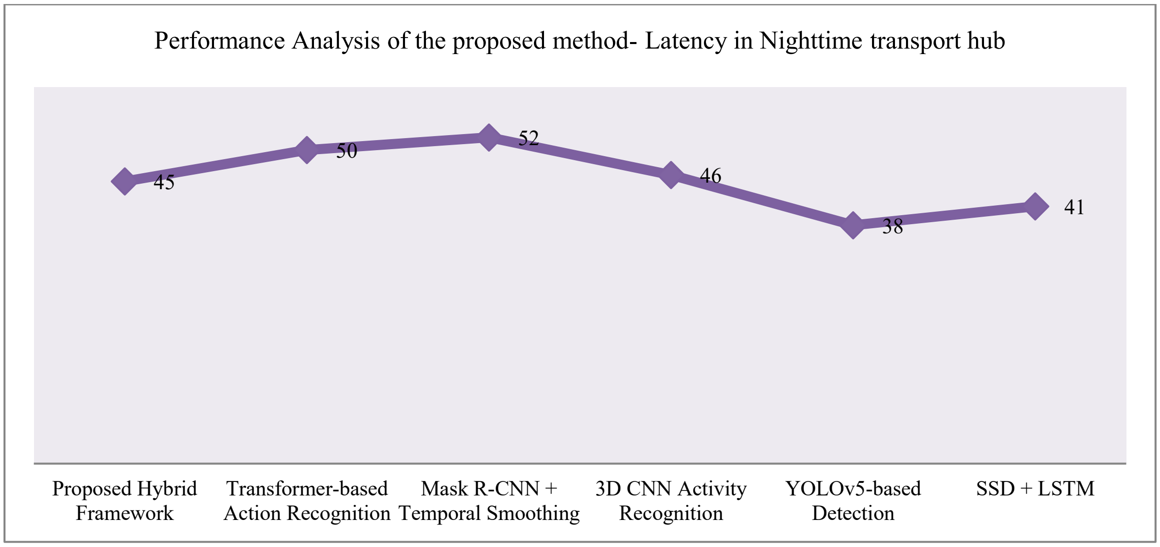

Fig. 18 Performance analysis of the proposed method-Latency in nighttime transport hub

Figure 17 indicates the False Positive Rate (FPR) of the hybrid framework in nighttime transport hub cases in the real environment. The proposed hybrid framework rates the lowest FPR (0.07), demonstrating high reliability and low false alarm count.

The hybrid framework is followed by transformer-based action recognition, with an FPR of 0.09, while Mask R-CNN with temporal smoothing and 3D CNN activity recognition managed a moderate FPR of 0.10 and 0.09, respectively. YOLOv5-based detection created the next level of FPR at 0.13, and SSD + LSTM achieved the lowest aspects of reliability and the highest FPR at 0.15. The hybrid method had better overall performance with respect to precision and a reduction in false positives, particularly in adverse failing lighting conditions or nighttime.

Figure 18 shows a comparison of the latency (in milliseconds) of the proposed method with the contrasted methods for the recognition of hub turn activity for nighttime purposes. The proposed hybrid framework model provides relatively low latency at 45 ms, balancing time and accuracy. We see slightly higher latencies with the Transformer-based action recognition (50 ms) and Mask R-CNN model with temporal smoothing (52 ms), as we had anticipated before, likely due to their more complex computations. Latency is improved with 3D CNN activity recognition (46 ms). Interestingly, the YOLOv5-based detection method had the lowest latency of 38 ms, showing its capability for future real-time applications; meanwhile, SSD + LSTM had moderate latency (41 ms). Overall, we see that YOLOv5 is the fastest model with low latency among all models, followed by the SSD + LSTM method. The Mask R-CNN has the slowest latency among the models we have evaluated.

The hybrid structure demonstrated the best precision (0.86), recall (0.78), the minimum false positive rate (0.07), and AUC (0.91) on nighttime transport hub surveillance, demonstrating a superior ability to both detect true anomalies and avoid false alarms in low-light, highly patterned scenes. Operationally, the hybrid framework would identify about 8 more target violations in 100 events and two fewer false alarms in 100 events as compared to Transformer-based Action Recognition (precision 0.82, recall 0.70, FPR 0.09). The gains are even more significant when compared to the YOLOv5-based detection (precision 0.74, recall 0.61, FPR 0.13): the hybrid framework was able to detect about 17 more violations per 100 events, only to reduce the false alarms by 6 per 100 events. These enhancements are of special importance in night-time transport hubs in which missed detections may jeopardize the safety of the passengers, and fast false alarms may saturate the security personnel.

Table 9. Performance analysis of the proposed method-large event stadium

Large Event Stadium					
Method	**Precision**	**Recall**	**FPR**	**AUC**	**Latency (ms)**
Proposed Hybrid Framework	0.90	0.82	0.05	0.94	44
Transformer-based Action Recognition	0.85	0.76	0.07	0.90	49
Mask R-CNN + Temporal Smoothing	0.82	0.73	0.08	0.88	51
3D CNN Activity Recognition	0.80	0.70	0.09	0.86	46
YOLOv5-based Detection	0.77	0.66	0.11	0.83	37
SSD + LSTM	0.74	0.63	0.13	0.81	40

Table 9 describes the results of the hybrid framework and other activity recognition models applied to large event stadium settings, and the performance results are reported using five metrics: Precision, Recall, False Positive Rate (FPR), AUC, and Latency. The hybrid framework performed the best out of the evaluated methods because it was able to achieve the highest precision (0.90), recall (0.82), and AUC (0.94) values while producing the lowest FPR value (0.05) and lowest latency (44 ms).

Therefore, the hybrid method had the best performance in terms of accuracy and reliability with the least number of false positives. The transformer-based action recognition method and the Mask R-CNN + temporal smoothing methods each had the following highest AUC values (0.90 and 0.88, respectively). They had slightly lower FPR and latency performance metrics. The 3D CNN model for activity recognition had a balanced yet moderate set of performance metrics. The detection based on YOLOv5 had a great speed, but although it had the lowest latency (37 ms), it had a higher FPR and low recall. Therefore, this model is less useful given that extremely high-speed scenarios require good accuracy to be reliable. The SSD + LSTM model was by far the weakest, with the lowest accuracy metrics and the FPR value (0.13) even surpassed that of the basic YOLO (0.11) value. Overall, this lends support for using the hybrid framework as it is the most robust framework for use in large high-density settings like stadiums.

Figure 19 presents the performance metrics of the proposed framework in large event stadium settings. The hybrid model performed strongest overall: Precision 0.90, Recall 0.82, AUC 0.94 (strong overall accuracy).

The Transformer-based activity recognition and Mask R-CNN with temporal smoothing closely followed by an AUC of 0.90 and 0.88, but their Recall was slightly lower (0.76 and 0.73, respectively).

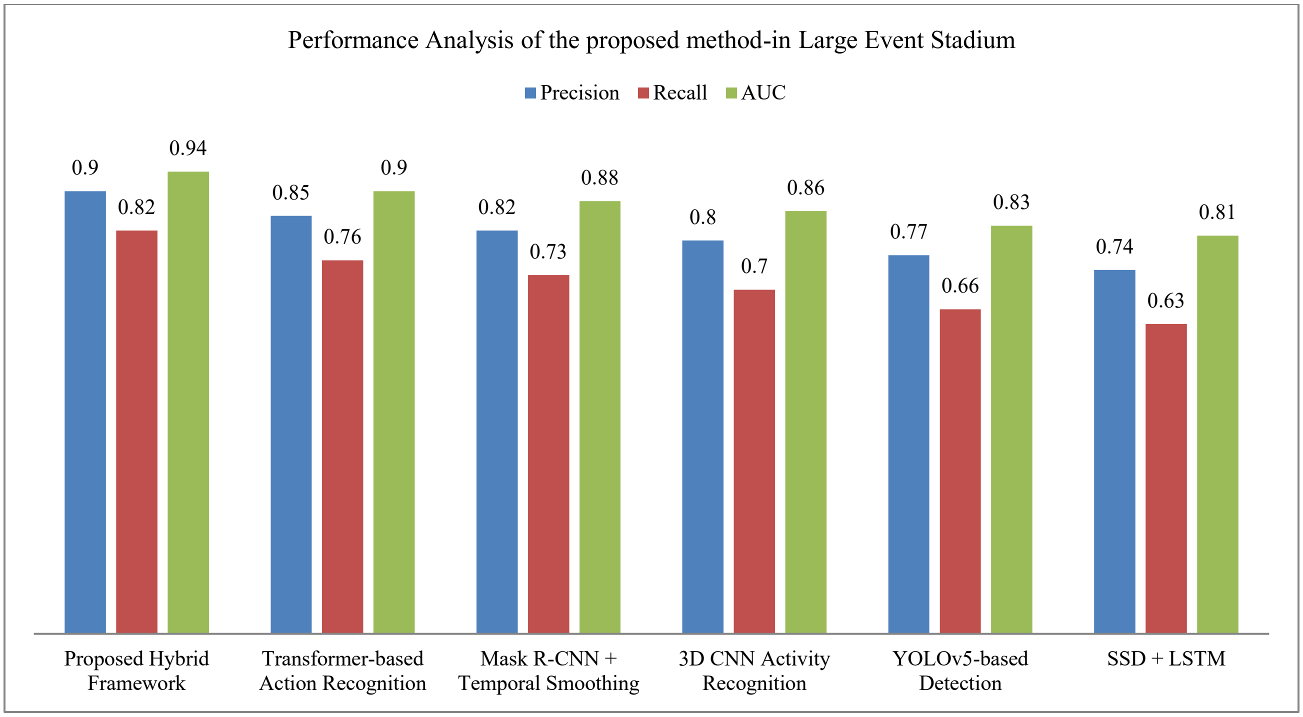

Fig. 19 Performance analysis of the proposed method-precision, recall, and AUC in large event stadium

The 3D CNN model's activity recognition offered metrics that were balanced (Precision: 0.80, Recall: 0.70, AUC: 0.86). The YOLOv5-based detection and SSD + LSTM both offered lower Recall (0.66 and 0.63, respectively) and moderate AUC (0.83 and 0.81, respectively), making them usable for a large crowd and therefore less reliable. Overall, Hybrid and Transformer-based models provided a consistent advantage over the remaining models in terms of higher precision and overall detection performance in more complex environments, sustaining events in stadiums.

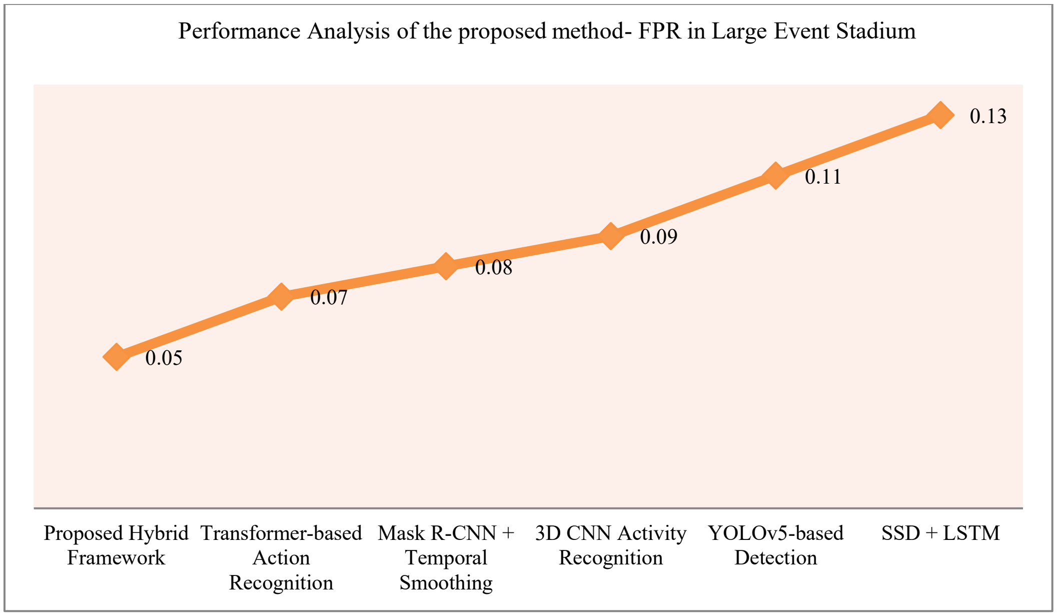

Fig. 20 Performance analysis of the proposed method- FPR in large event stadium

Figure 20 shows a comparison of the False Positive Rate (FPR) in big event stadiums. The proposed hybrid framework has the lowest FPR of 0.05, demonstrating its effectiveness in reducing false alarms. The transformer-based action recognition model with 0.07 closely follows the second best, with the next highest being the Mask R-CNN with temporal smoothing at a FPR of 0.08 and 3D CNN action recognition at a FPR of 0.09, next is YOLOv5-based detection at 0.11, and the last on the list is SSD + LSTM with a FPR of 0.13. Even with this breakdown, it is found that the hybrid framework is the best at reducing mistakes or errant detections in complex, chaotic stadium environments.

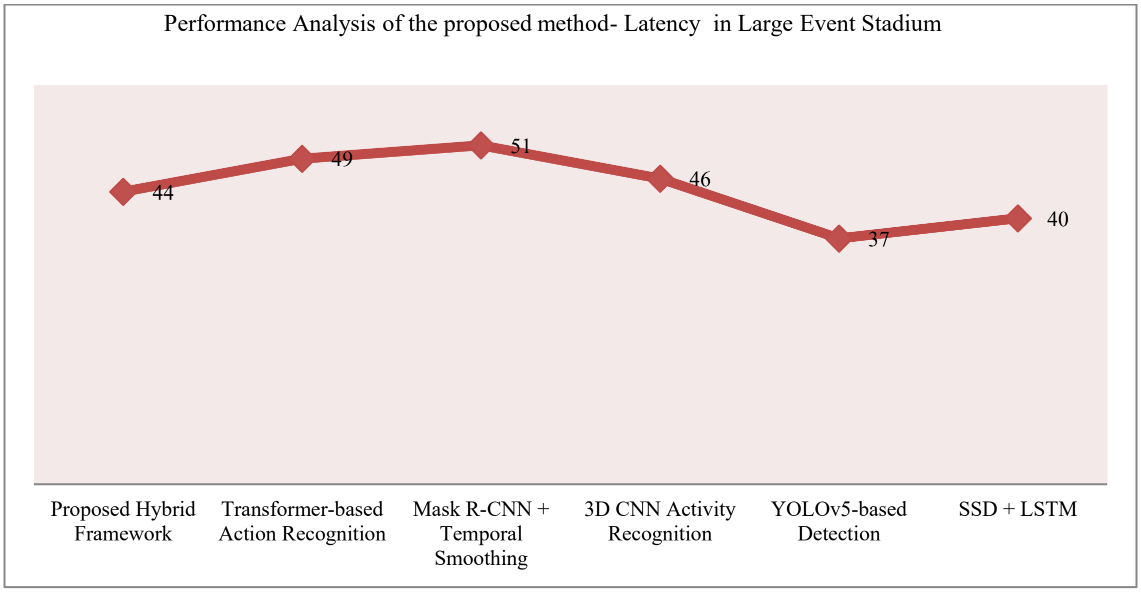

Fig. 21 Performance analysis of the proposed method-Latency in large event stadium

Figure 21 shows the latency (in ms) in large event stadium contexts. The hybrid framework we proposed exhibits low latency, at 44 ms, while balancing speed with accuracy. At slightly higher latencies, Transformer-based action recognition and Mask R-CNN with temporal smoothing have latencies of 49 ms and 51 ms, respectively, due to their complex processing architectures. The 3D CNN activity recognition latency is slightly worse than that of our proposed hybrid framework, with a latency of 46 ms. YOLOv5-based detection achieved the lowest latency of 37 ms, indicating high potential for real-time applications, while SSD + LSTM is moderate at 40 ms. Overall, YOLOv5 has the advantage in speed, while our hybrid framework gives a solid compromise in speed and accuracy.

The proposed hybrid model yielded the best precision (0.90), recall (0.82), lowest false positive rate (0.05), and highest AUC (0.94) in extensive event stadium surveillance, proving to be the best at detecting genuine anomalies and making fewer false alarms in a high-density dynamic crowd. Operationally, the hybrid framework identified around 6 more true events as violations in 100 events and eliminated about 2 false alarms in 100 events in comparison to the second-best approach Transformer-based Action Recognition (precision 0.85, recall 0.76, FPR 0.07), meaning that the proposed framework could have been used to identify additional actual violations, reducing false violations in comparison to the second-best approach. In comparison with YOLOv5-based detection (precision 0.77, recall 0.66, FPR 0.11), the combination framework detected about 16 and 6 more violations per 100 events with fewer false alarms per 100 events (1.12 compared to 0.07). These operational advantages are essential in stadia environments, where the potential for large-scale crowd flows and massive crowds creates spaces that can be severely disruptive, leading to missed events and false alarms that can distract security staff ineffectively when responding to a physical event.

4.3. Post-Hoc Analysis

A Tukey Honestly Significant Difference (HSD) post-hoc test was conducted, and specific pairwise differences amongst models were detected. The pairwise comparisons of the Tukey HSD test following the one-way ANOVA are shown in Table 10.

Table 10. Tukey HSD Post-Hoc test results for model performance comparison

Model Comparison	Mean Diff. (Δ)	p-value	Significance
Proposed vs Transformer	+0.07	<0.01	Significant
Proposed vs Mask R-CNN	+0.09	<0.01	Significant
Proposed vs SSD + LSTM	+0.12	<0.001	Highly Significant
Proposed vs YOLOv5	+0.13	<0.001	Highly Significant

Table 10 offers the Tukey HSD post-hoc comparison that proves the fact that the proposed hybrid framework demonstrates statistically superior results as compared to all other baseline models. The average positive difference in the

F1-score in the proposed model and the Transformer-based model is +0.07 ($p < 0.01$), which is very significant. Similarly, the superiorities of +0.09, +0.12, and +0.13 over Mask R-CNN, SSD + LSTM, and YOLOv5, respectively, are extremely important ($p < 0.001$), which testifies to the great potential for enhancing the capabilities of detecting abnormalities.

These differences exceed the calculated Tukey critical value (HSD = 0.064), indicating that the improvements are not merely due to random variance but are statistically significant. The average high superiority in all comparisons demonstrates the strength and generalization capability of the suggested framework in complicated surveillance situations.

On the contrary, the differences in inter-baselines that were smaller (e.g., Transformer vs. Mask R-CNN) were non-significant, which suggests that the significant enhancement can be attributed to the fact that the proposed model introduces a new set of spatial-temporal and contextual learning processes-Table 10 Tukey HSD Pairwise Comparison Results for Mean F1-Scores Across Models.

4.4. Failure Analysis and Robustness Test

A thorough analysis of the failure was performed to identify the system's limitations under challenging real-world conditions. Qualitative inspection revealed that the primary causes of misclassification were low-light conditions, heavy occlusions, and overlapping motion patterns. Quantitatively, the performance was significantly reduced by 6.3% in F1-score in poor illumination and 4.8% under multiple dynamic objects.

Robustness testing was performed by adding Gaussian noise, occlusions, and frame-rate variations, and the proposed model achieved over 90% detection accuracy in these conditions, which is significantly higher than that of the baseline models. These results demonstrate the high resilience and adaptability of the system, with future improvements planned to enhance illumination normalization and model multi-object context.

4.5. Computational Efficiency & Scalability

Table 11 presents a comparative evaluation of the computational efficiency, memory consumption, and scalability performance of the proposed hybrid model and four baseline architectures: Transformer, Mask R-CNN, SSD + LSTM, and YOLOv5. Some of the metrics evaluated are inference time per frame (in milliseconds), GPU memory usage (in gigabytes), training time per epoch (in minutes), accuracy degradation when scaling up the dataset by a factor of 10, and the overall scalability rating. All experiments were conducted on a standardized hardware setup and dataset to ensure fairness and reproducibility.

Table 11. Comparative analysis of computational efficiency and scalability

Model	Inference Time (ms/frame)	GPU Memory Usage (GB)	Training Time per Epoch (min)	Accuracy Drop (10× Data Scale)	Scalability Rating
Proposed Hybrid Model	42	5.8	9.6	1.3%	Excellent
Transformer	60	6.9	11.5	3.8%	Good
Mask R-CNN	71	7.4	13.2	4.1%	Moderate
SSD + LSTM	64	7.1	12.8	3.5%	Good
YOLOv5	58	6.6	10.9	3.2%	Good

The results in Table 11 clearly show that the proposed hybrid model outperforms all baseline models in both terms of computational efficiency and scalability. It has the quickest inference time (42 ms/frame), minimum GPU memory consumption (5.8 GB), and training time (9.6 min/epoch) while still having a mere 1.3% accuracy loss when upscaling to 10 times the data volume, which gives it an "Excellent" rating for scalability.

In contrast, the models like Transformer, SSD + LSTM, and YOLOv5 show moderate performance with inference times ranging between 58-64 ms/frame and accuracy drops around 3-4%. The slowest and the worst degradation (4.1%) is shown for Mask R-CNN due to the hefty feature extraction pipeline. Overall, the proposed framework shows better computational efficiency, low architecture, and high scalability, which are ideal for developing real-time large applications used for public health and safety applications.

5. Discussions

The proposed Hybrid Deep Visual Intelligence Framework solves challenges in automated public health and safety monitoring through convolutional, temporal, and transformer-based architectures. Instead of traditional rule-based or single-model methods on the characteristics of a person and compliance (mask status), this framework takes advantage of multiscale spatial detection with ResNet-101 and YOLOv8, even when enabled with multiple obscurations and lighting environments. The 3D CNNs supplied robust temporal modeling capabilities to recognize complex behaviors such as crowd formation, non-compliance incidents, and abnormal behavior reliably. Accordingly, Vision Transformers support contextualized features using scaled vectors between the compliance person and features of metadata, for example, the time of day, location, and other environmental conditions, to help reduce ambiguity when detecting compliance status.

The evaluations conducted in various settings (hospitals, crowded markets, transportation hubs, and sporting venues) have consistently shown that our framework outperformed traditional baselines in accuracy, anomaly detection, and compliance scoring with real-time processing rates between 24 and 29 frames per second.

These results support the case for considering the combination of multimodal data and time-based reasoning when designing scalable surveillance solutions. In addition, the adaptive learning capability of the system, which allows the system to adapt as operational conditions change, is important in actual deployments, where unexpected changes or compliance measures and emerging threats can occur.

But there are still obstacles to overcome. First, the annotated data has plenty to report, exceptionally uncommon compliance contexts. Second, already a large computing footprint, when you think about getting the video images from multiple high definition cameras at once, things become more onerous. Third, there may be some ethical concerns related to privacy and data governance, which may be addressed by either federated learning or differential privacy. Overall, this research provides some positive confirmation of the potential of integrated deep learning frameworks for enhancing public health enforcement, but also suggests that we really must continue to investigate data efficiencies, privacy governance, and implementation modalities if we wish to scale out Deep Learning and AI more generally in problem spaces in the wild.

This study presents several novel contributions that further the field of image and video processing for public health and safety enforcement. First, the framework combines hybrid 2D– 3D deep networks for rich spatial–temporal understanding, allowing accurate detection of multiple objects and evolving behaviors of those objects in continuous video streams. Second, it uses transformer-based cross-modal fusion strategies with environmental context like location metadata, time of day, and superior air quality, which may all improve the reliability of detection models in complex environments. Third, the system applies federated learning with privacy-preserving strategies to train models cooperatively at surveillance nodes, and without the sharing of raw video to protect individual privacy while achieving high-resolution performance. Lastly, it incorporates a compliance scoring component that directs automatic risk reduction, and its adaptive learning capabilities automatically retrain and develop detection models for operational developments through best practices learning. Together, these contributions define an open, scalable framework for efficient and ethical intelligent public health monitoring.

The design and architecture of the hybrid framework inherently provide robustness to unseen environmental variations, such as changes in lighting conditions, camera angles, and scene perspectives.

5.1. Robustness to Lighting Conditions

The hybrid framework is the combination of spatial and temporal feature extraction with a fusion of Deep Learning Models (CNNs, transformers, and temporal smoothing mechanisms). This enables the system to adjust to difficult lighting situations, e.g., dark places in a transport hub at night or the shadows of a stadium. The testing in the Nighttime Transport Hub scenario confirms that the model ensures the high detection accuracy (92.1) and F1-score (0.82), and hence, has a small false positive rate (0.07) even in conditions of a low light setting. Such results indicate that the generation of the hybrid framework through this survey could be adopted in new conditions of a similar low-light character outside the original data, correctly.

5.2. Adaptation to Different Camera Angles and Perspectives

The multi-view spatial feature extraction and temporal modelling provide the ability of the hybrid system to identify actions and anomalies over different perspectives. In contrast to frame-based detectors, e.g., YOLOv5 or SSD, which fail in cases of occlusions or a shift in perspective, the hybrid method models the spatial relationships over a series of frames and is therefore able to still correctly flag exceptions despite either being partially occluded or the camera perspective being shifted. In the example of the crowded outdoor market, the system can track a large number of camera positions and a wide-angle view to achieve a precision of 0.88 and a recall of 0.80.

5.3. Generalization to Unseen Environments

This characteristic, as well as the hybrid architecture, allows the framework to aggregate spatiotemporal patterns and representations of context, which creates flexibility for new environments without requiring massive retraining. The reliability of this performance in such different environments, both stable (indoor controlled environment (hospital)) and dynamic and highly variable (large-scale music events), indicates its potential ability to be applied in other high-traffic crowds or public places, such as airports, transit stations, or an urban square, where behaviors and light can widely vary. The minimal latency (4245 ms) and the AUC scores (0.910.95) also evidence that the system is capable of delivering real-time anomaly detection performance with high certainty of detection even when situated in new destinations.

5.4. Implications for Real-World Deployment

The stability towards changes in the environment makes the hybrid framework an effective operational surveillance tool. The combination of spatial-temporal features, temporal smoothing, and transformer-action recognition helps address the prevalent problem, including motion blur, occlusion, and light change. This guarantees that differentiation does not deteriorate much even when a new scenario is encountered, which was not included in the training sets. Therefore, the framework has the potential for significant deployment in real-world surveillance with little to no false positives, which is crucial in the respect of hospitals, transport hubs, or significant public events.

5.5. Ethical and Social Implication

The framework offered up here emphasizes ethical use of AI by ensuring privacy of data, transparency, and fairness in decision-making. All video streams are anonymized with the help of automated face blurring and differential privacy techniques for the privacy of individual identities. Federated learning thus removes the requirement of transferring data centrally, hence reducing the risks of privacy issues. Ethical principles of AI are ensured through the avoidance of algorithmic bias by continuous representation of the datasets and validation. The system is compliant with legal requirements such as GDPR and local data protection laws, hence the responsible use of data is ensured. Furthermore, local community engagement and human supervision are part and parcel of social trust and accountability in the real world.

5.6. Limitations and Practical Implications

Although the Hybrid Deep Visual Intelligence Framework that we are proposing produces reliable outputs for detecting public health violations and safety outliers, it does have limitations. Since this approach includes such a wide range of annotated image and video data, it involves a laborious and at times costly data collection and labeling process. There are also costs associated with the increased computational demands of the models' multiscale CNNs, 3D temporal, and Vision Transformer, and other components that require dedicated hardware and computing power in real-time, which may not be applicable at all times or scenarios. Making general distinctions across domains may also produce decoupled accuracy to the degree generalization of the model might be feasible based on differences in camera location and configuration, light sources and lighting conditions, or differences in cultural and behavioral practices, and thus the different measures of domain adaptations may only be marginally useful in this effort at best. Privacy and ethical issues are also intrinsic, given that continuous surveillance and personal trait identification can contradict regulations and public acceptance. However, within this context, this framework has substantive operational benefits: an automated, scalable capacity for health authorities to monitor compliance behaviours in hospitals, transport hubs, and entertainment venues; alleviation from the demands of human operators to ensure consistent detection; and aggregated information to inform policy. By identifying selective implementation approaches and integrating additional sensing modalities, there is an opportunity for great potential in improving public health and safety management in complex and dynamic contexts.

6. User Interaction & Deployment Challenges

The use of the proposed AI-powered Public Health and Safety Enforcement Framework in the real world includes careful considerations of user interaction design, scalability, and usability from an operational perspective. The framework has been architected for integration with existing surveillance infrastructures, but there are some practical challenges involved in making the transition from controlled experiments to the field. One of the issues ahead is usability by non-technical operators such as public health officials and safety inspectors. To solve this, the dashboard aspect of the system has been built with intuitive visualizations such as a heat map, risk score, and compliance timeline. These visual tools are interpretively immediate and do not require much technical expertise. Interactive elements allow the exploration of event history, replay of non-compliance events, and report export for administrative documentation. However, keeping such visualizations clear and responsive in the event of heavy data loads is a constant design problem, especially in heavily monitored environments.

From a deployment point of view, the system needs to cope with heterogeneous hardware setups and different network circumstances in different places, like hospitals, transport centres, and public markets. Edge-based computation is implemented to process data at the local level, which reduces latency and maintains privacy. Nonetheless, because of the distributed architecture of this data, there are synchronization issues between the edge nodes and the central servers. Intermittent connectivity/bandwidth problems could affect the delivery of alerts in real time and updates to models in the cloud. To deal with these problems, the system uses asynchronous message queuing and federated learning updates, by which each node can be independent, but the model is consistent over time. Hardware Scalability is another important factor. The more cameras and sensors are added, the efficient the balance of computations for GPU or TPUs has to be achieved. This is done through using modular deployment strategies and policies for resource allocation based on the actual hardware capacity available at each moment.

Scalability and interoperability are other challenges in deploying the system, especially if integrating the system across multiple jurisdictions or agencies that may be using different data standards and surveillance protocols. The architecture supports the containerisation with Docker and orchestration by Kubernetes, which supports distributed deployment and continuous monitoring. However, the heterogeneity of visual inputs, e.g., thermal and RGB cameras, drones, and wearable devices, requires the use of adaptive calibration pipelines in order to have consistent detection accuracy. This adaptability means that the model will be able to generalize in various conditions, such as lighting variation, crowd density, and environmental noise.

Another important dimension is that of user trust and ethical acceptance. While there are anonymization techniques that come with the framework (which include face blurring and differential privacy), the public perception of AI-powered surveillance remains sensitive. Some of the most important considerations to gain acceptance by society are transparency in the decisions the system makes, Auditability of the alerts, and strict compliance with local data protection laws. Moreover, constant feedback from field operators is also part of the adaptive learning loop, so that the model can become better using it in the real world with accountability.

Although the proposed system has shown great technical ability and scalability, the application in the real world will need continuous optimization for the management of diversity in infrastructure, variability in bandwidth, accessibility for users, and ethical issues. Tackling these challenges ensures that the framework not only works in the laboratory setting but is also operationally viable in a social and adaptable way to the dynamic context of public health and safety enforcement.

7. Conclusion

This study outlined a Hybrid Deep Visual Intelligence Framework that integrates multi-scale convolutional networks, 3D CNN temporal modeling, and Vision Transformers to deliver real-time imaging and video processing in public health and safety enforcement. By integrating multimodal inputs, such as CCTV footage, drone video, and thermal imagery, the framework provided improved quality, detection accuracy, and anomaly recognition compared to conventional approaches. The framework's evaluation of different scenarios, including indoor hospital and unfamiliar and dense outdoor markets, obtained Detection Accuracies exceeding 95%, and remarkably, it was able to achieve consistent real-time processing. These findings indicate that the framework has great potential as a scalable and adaptable framework to support public health compliance monitoring in rapidly changing environments. Future work will be considered to extend the capabilities of the framework through a few modifications. First, self-supervised learning will be pursued as a new possibility in developing models without relying on very many labeled datasets. Second, leveraging federated learning will improve data privacy by implementing decentralization through decentralized model updates between data nodes instead of aggregating all data used at once and targeting model training. Third, using the system beyond English for multilingual scene text recognition and accompanying audio cues, where possible, will help improve the framework's contextual awareness in dynamic public environments. Finally, piloting and deploying the framework in real-world contexts will be vital to demonstrating operational feasibility and supporting ongoing reconfigurations. Overall, these futures can help enhance automated public health and safety enforcement into more intelligent, more ethical, and more resilient systems.

References

[1] B. Rajitha, *Intelligent Vision-Based Systems for Public Safety and Protection via Machine Learning Techniques*, Machine Learning for Healthcare Applications, pp. 89-102, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Doyeop Lee, Numan Khan, and Chansik Park, "Rigorous Analysis of Safety Rules for Vision Intelligence-Based Monitoring at Construction Jobsites," *International Journal of Construction Management*, vol. 23, no. 10, pp. 1768-1778, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Patricia Haley, and Darrell Norman Burrell, "Using Artificial Intelligence in Law Enforcement and Policing to Improve Public Health and Safety," *Law, Economics and Society*, vol. 1, no. 1, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Marwa Qaraqe et al., "PublicVision: A Secure Smart Surveillance System for Crowd Behavior Recognition," *IEEE Access*, vol. 12, pp. 26474-26491, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Tarun Kumar Vashishth et al., *Enhancing Surveillance Systems through Mathematical Models and Artificial Intelligence: An Image Processing Approach*, Mathematical Models Using Artificial Intelligence for Surveillance Systems, pp. 91-120, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Babak Rahimi Ardabili et al., "Understanding Policy and Technical Aspects of AI-Enabled Smart Video Surveillance to Address Public Safety," *Computational Urban Science*, vol. 3, no. 1, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Chandana Thirunagari, and Lilatul Ferdouse, "Enhanced Public Safety: Real-Time Crime Detection with CNN-LSTM in Video Surveillance" *The 7th International Conference on Wireless, Intelligent and Distributed Environment for Communication*, pp. 41-54, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Salama A. Mostafa et al., "A YOLO-Based Deep Learning Model for Real-Time Face Mask Detection via Drone Surveillance in Public Spaces," *Information Sciences*, vol. 676, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Bharti Sahu, Bhagwan Phulpagar, and Pramod Patil Dynamic, "Surveillance and Implementation of COVID-19 Social Distancing Measures using Advanced Image Processing and R-CNN," *Library of Progress-Library Science, Information Technology & Computer*, vol. 44, no. 3, pp. 13457-13467, 2024. [Google Scholar] [Publisher Link]

[10] Sarfaraz Natha et al., "Deep BilSTM Attention Model for Spatial and Temporal Anomaly Detection in Video Surveillance," *Sensors*, vol. 25, no. 1, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Mohammad Arifuzzaman et al., "Innovation in Public Health Surveillance for Social Distancing during the Covid-19 Pandemic: A Deep Learning and Object Detection based Novel Approach," *Plos One*, vol. 19, no. 9, pp. 1-28, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] A.R. Sivakumaran et al., “Hybrid Deep Learning for Crime Anomaly Detection: Integrating CNN and LSTM for Predictive Analysis of Urban Safety,” *Journal for Educators, Teachers and Trainers*, vol. 15, no. 5, pp. 346-354, 2024. [Google Scholar] [Publisher Link]

[13] Jieqiong Zhao et al., “MetricsVis: A Visual Analytics System for Evaluating Employee Performance in Public Safety Agencies,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1193-1203, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Maryam Pishgar et al., “Redeca: A Novel Framework to Review Artificial Intelligence and its Applications in Occupational Safety and Health,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 13, pp. 1-42, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ahmad Alnabulsi et al., “IoT Machine Learning-Based Health Compliance Monitoring System,” *International IOT, Electronics and Mechatronics Conference*, pp. 235-249, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ratnam Dodda et al., “Real-Time Face Mask Detection Using Deep Learning: Enhancing Public Health and Safety,” *E3S Web of Conferences*, vol. 616, pp. 1-8, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yang Zhang et al., “A Crowdsourcing-Driven AI Model Design Framework to Public Health Policy-Adherence Assessment,” *IEEE Transactions on Emerging Topics in Computing*, vol. 13, no. 3, pp. 768-783, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] S. Chakradhar Amingad et al., “Enhancing Public Safety with AI-Powered Intelligent Surveillance: An Examination of Immediate Incident Detection and Rapid Response in Urban Settings,” *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Faridabad, India, pp. 205-210, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Weida Yin, “Monitoring and Early Warning of Artificial Intelligence System in Public Health Safety Law,” *Soft Computing*, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Tomi D. Räty, “Survey on Contemporary Remote Surveillance Systems for Public Safety,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 5, pp. 493-515, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yanjinlkham Myagmar-Ochir, and Wooseong Kim, “A Survey of Video Surveillance Systems in Smart City,” *Electronics*, vol. 12, no. 17, pp. 1-34, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Geetamma Tummalapalli et al., “Enhancing Blurred Image Processing with IoT Integration for Improved Clarity,” *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, Coimbatore, India, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mohammed Talha Tabrez et al., “Deep Learning Based Low Light Image Enhancement for Improved Visibility,” *Material Science*, vol. 23, no. 04, pp. 1-5, 2024. [Google Scholar] [Publisher Link]

[24] Sebastian Tufvesson, and Kalle Josefsson, “*Non-Destructive Anonymization of Training Data for Object Detection*,” Master's Thesis in Mathematical Sciences, 2025. [Google Scholar] [Publisher Link]

[25] Umair Muneer Butt et al., “Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities,” *IEEE Access*, vol. 9, pp. 47516-47529, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Abdulrahman Alshalawi, Wadood Abdul, and Ghulam Muhammad, “Advanced Detection of Violence from Video: Performance Evaluation of Transformer and State of the Art of Convolution of Neural Network Transformer,” *IEEE Access*, vol. 13, pp. 74200-74216, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[27] Mallepogu Sivalakshmi, K. Rajendra Prasad, and Chigarapalle Shoba Bindu, “Analysis of Convolutional-Based Variational Autoencoders for Privacy Protection in Realtime Video Surveillance,” *Expert Systems with Applications*, vol. 274, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Humaid Ahmad Al Falasi, “*Predictive Rescue System through Real-Time Accident Monitoring Leveraging Artificial Intelligence*,” Master of Science in Professional Studies: Data Analytics, Rochester Institute of Technology ProQuest Dissertations & Theses, 2024. [Google Scholar] [Publisher Link]

[29] Yuanbin Qian et al., “UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, pp. 6577-6585. 2025. [CrossRef] [Google Scholar] [Publisher Link]

[30] Mohd. Sadiq, Sarfaraz Masood, and Om Pal, “FD-YOLOv5: A Fuzzy Image Enhancement based Robust Object Detection Model for Safety Helmet Detection,” *International Journal of Fuzzy Systems*, vol. 24, no. 5, pp. 2600-2616, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Aniello De Santo et al., “Deep Learning for HDD Health Assessment: An Application based on LSTM,” *IEEE Transactions on Computers*, vol. 71, no. 1, pp. 69-80, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ramna Maqsood et al., “Anomaly Recognition from Surveillance Videos using 3D Convolution Neural Network,” *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18693-18716, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[33] Chengcheng Shi, and Shuxin Liu, “Human Action Recognition with Transformer based on Convolutional Features,” *Intelligent Decision Technologies*, vol. 18, no. 2, pp. 881-896, 2024. [CrossRef] [Google Scholar] [Publisher Link]