GMR

The Original

# AN INTELLIGENT DEEP LEARNING FRAMEWORK USING YOLO-BASED CNN–LSTM–ATTENTION NETWORKS FOR BIOMEDICAL IMAGE ANALYSIS AND MOLECULAR PATTERN RECOGNITION

**[1*]Karthick Manoharan, [2]Kannan N**

[1*]Research scholar, Department of Computer Science, School of Engineering Technology, CMR University, Bengaluru, Karnataka, India - 560043.
[2]Professor, Department of Computer Science, School of Engineering Technology, CMR University, Bengaluru, Karnataka, India - 560043.
*Corresponding Author: Karthik5manoharan@gmail.com

## ABSTRACT

**Background**: The significance of intelligent monitoring systems in railway safety has grown, particularly in detecting animal intrusions and dangerous objects on tracks. Advances in deep learning, computer vision, and biomedical image intelligence have facilitated automated biological pattern recognition and large-scale image analysis, paving the way for developing effective surveillance systems that enhance public safety and environmental monitoring.
**Objectives**: This research introduces an intelligent deep learning framework that combines YOLO-based object localization with a CNN–LSTM–Attention architecture for automated railway surveillance, detection of animal intrusions, recognition of hazardous objects, and analysis of biological patterns. The framework utilizes advanced image preprocessing methods, such as image enhancement, noise reduction, normalization, and region-of-interest (ROI) extraction, to enhance visual information quality across various environmental conditions.
**Methods**: The YOLO detection module precisely identifies animals, hazardous objects, and railway track areas from surveillance video sequences. The convolutional neural network extracts distinct spatial features, while the long short-term memory network captures temporal relationships from successive video frames. Additionally, the attention mechanism selectively highlights biologically significant regions and hazardous objects, enhancing classification accuracy and decision-making in intelligent railway monitoring.
**Results**: The proposed framework uses computational feature optimization and statistical learning techniques to effectively represent high-dimensional visual features. Experimental results show superior performance in accuracy, precision, recall, F1-score, and robustness compared to traditional deep learning methods. The system efficiently detects track intrusions, hazardous objects, and animal movements, reducing false alarms and improving real-time surveillance capabilities.
**Conclusion**: The artificial intelligence-driven framework offers an efficient solution for intelligent railway surveillance, detection of hazardous objects, monitoring of animal intrusions, and biological image intelligence. It supports future applications in intelligent transportation systems, bio-inspired computer vision, automated safety monitoring, environmental surveillance, and next-generation smart railway infrastructure.

**KEYWORDS**: Deep Learning, Railway Safety, Animal Intrusion Detection, Hazardous Object Detection, YOLO Detection, CNN–LSTM–Attention, Biological Pattern Recognition, Intelligent Surveillance, Computer Vision, Intelligent Transportation Systems.

## INTRODUCTION

Safety and security have turned into a tough challenge these days, and unauthorized access and trespassing can be seen in the railway environment. The current surveillance systems lack the capability to detect and predict in real-time because of their reliance on manual surveillance and lack of static image analysis capability. To overcome these challenges, this research presents an integrated framework that is proposed for smart railway surveillance combining a Convolutional Neural Network (CNN) and a multimedia preprocessing pipeline, which is based on temporal analysis. The time sequence analysis along with state of the art extraction of visual frames spatial features of the proposed system enhances the real-time detection accuracy of suspicious events and intrusions. This method will increase detection rate by tracking the movement pattern with time and also greatly reduce the false alarm rate, allowing an automated response system and smart monitoring and preventive safety system for railway systems. This system should be created to help transition to an all-AI-operated system that would also help with the modernization of railroads. A smart safety system is required for the high-speed railway system to avoid accident from obstacle.

However, the research aims to reduce the number of accidents and enhance railway safety. A deep classifier network to identify obstacles on railway tracks together with two-dimensional Singular Spectrum Analysis (2D-SSA) for image decomposition [1]. The integration of deep learning with 2D-SSA enhances obstacle detection. The application of an Artificial Intelligence-based Safety and Security System for Rail Crossing Traffic (AISS4RCT) to improve safety and security at railway crossings [2]. The system utilizes deep learning and image processing to automatically recognize dangerous situations in real time based on the presence of vehicles, pedestrians, movement patterns, and traffic signs. Meanwhile, railways are open systems that are vulnerable to terrorism, trespassing, and suicide. Recent machine learning-based computer vision techniques for CCTV surveillance in railway environments, highlighting methods for static object detection, human tracking, threat prevention, behavior recognition, and privacy-aware surveillance [3].

Accordingly, The YOLOv3 model could detect railway intrusions under challenging conditions such as rain and darkness with an overall detection accuracy exceeding 90%, outperforming manual monitoring [4]. The proposed system generated notifications containing intrusion time, location, object class, object size, and captured images, thereby improving accident prevention and railway safety monitoring. Railways continue to face threats from terrorism and vandalism, making intelligent CCTV surveillance increasingly important. The rapid evolution of digital video surveillance technologies and modern video compression standards that provide high-quality video with reduced bandwidth and storage requirements [5]. Furthermore, that contemporary digital recording systems incorporate intelligent video analytics such as face recognition, unattended object detection, and intrusion detection [5]. The study also emphasized important design considerations for large-scale railway surveillance systems, including image resolution, frame rate, storage duration, redundancy, and digital video recorder (DVR) configurations.

Railway stations represent critical components of transportation infrastructure and require continuous safety monitoring. An artificial intelligence framework based on computer vision and convolutional neural networks to identify fall, slip, and trip events for proactive risk management [6]. The framework significantly improved motion analysis accuracy and enabled timely detection of hazardous situations. Similarly, An AI-driven surveillance architecture that combines computer vision and artificial intelligence to automate the monitoring of multiple surveillance cameras simultaneously [7]. The scalable framework demonstrated its effectiveness in railway station surveillance while reducing the burden on human operators.

To improve operational reliability without compromising safety certification, separating intelligent monitoring functions from safety-critical railway systems and highlighted the application of deep learning for predictive maintenance using smart audio and video sensors [8]. Dangerous passenger activities during train travel and proposed an AI-enabled surveillance framework for real-time safety monitoring [9]. In addition, A low-visibility obstacle detection system integrating infrared sensors, ultrasonic sensors, computer vision, and convolutional neural networks [10]. The system employed an ESP32 controller with GPS, GSM, LoRa, Wi-Fi, and cloud-based monitoring through ThingSpeak for real-time obstacle localization and communication. Likewise, Artificial intelligence can improve railway safety by detecting suspicious objects, monitoring abnormal human activities, and identifying obstacles along railway tracks [11]. Furthermore, Ooppakaew et al. (2025) integrated ResNet-50 with semantic segmentation models including SegNet, U-Net, FCN-8, FCN-16, and FCN-32 for railway track detection, demonstrating that optimized learning rates significantly enhanced segmentation accuracy using a dataset of 5,000 images.

Although video surveillance-based railway intrusion detection has gained considerable attention, its effectiveness remains limited by insufficient training samples and the presence of complex background environments [12]. Railway wheel fault diagnosis models using performance metrics such as Accuracy, Precision, Recall, Miss Rate, and Area Under the Curve (AUC) [13]. Their findings indicated that EfficientNet-B7 outperformed Xception in terms of recall, miss rate, and AUC while maintaining an overall accuracy of 91.1%, making it suitable for railway rolling stock wheel fault diagnosis and predictive maintenance. The absence of reliable obstacle detection systems significantly increases the likelihood of railway accidents [14]. An obstacle detection model based on SSD-MobileNet and deep neural networks deployed on a Raspberry Pi platform [15]. The system accurately detected and classified multiple objects within a predefined Region of Interest (ROI) using both static images and real-time video streams.

Landslide monitoring systems utilizing wireless sensor networks offer ongoing real-time data on slope conditions and environmental factors [16]. Their practical application underscored the sensor networks' effectiveness in early landslide detection, while also revealing issues related to sensor dependability, communication, and prolonged field operations. An early landslide detection model enabled by the Internet of Things (IoT), aiming for minimal power usage, rapid processing, and energy-efficient monitoring [17]. Their experiments showed enhanced detection efficiency alongside reduced energy use for continuous environmental monitoring. An IoT-based system for landslide monitoring and detection, which combines various environmental sensors for real-time data collection and hazard identification [18]. The framework they proposed showed dependable monitoring performance and facilitated timely warnings under diverse environmental conditions. The LANDSLIDE MONITOR system, which integrates environmental sensing with continuous data collection for real-time landslide monitoring and hazard evaluation [19]. Their method improved early warning capabilities and bolstered disaster readiness by promptly identifying hazardous events. An IoT-driven microseismic sensing platform for intelligent landslide detection, demonstrating that combining microseismic signals with IoT-based monitoring greatly enhanced hazard detection accuracy, system reliability, and continuous environmental monitoring for disaster risk management [20].

## METHODOLOGY

The proposed smart railway surveillance system is designed in the form of a high-end-to-end visual-temporal intelligence pipeline, which will enable real-time detection of trespassing events and hazardous objects correctly in the railway scenes. The suggested approach starts from the real-time processing of railway video streams acquired from locomotive mounted cameras to maintain the consistency of the video in the pilot's view. Video frames are then preprocessed in a structured way for these video streams: They are decoded at a fixed frame rate (25 FPS), resized to a fixed size and post-processed (noise reduction, contrast normalization) to enhance the visual quality of the video in different lighting situations.

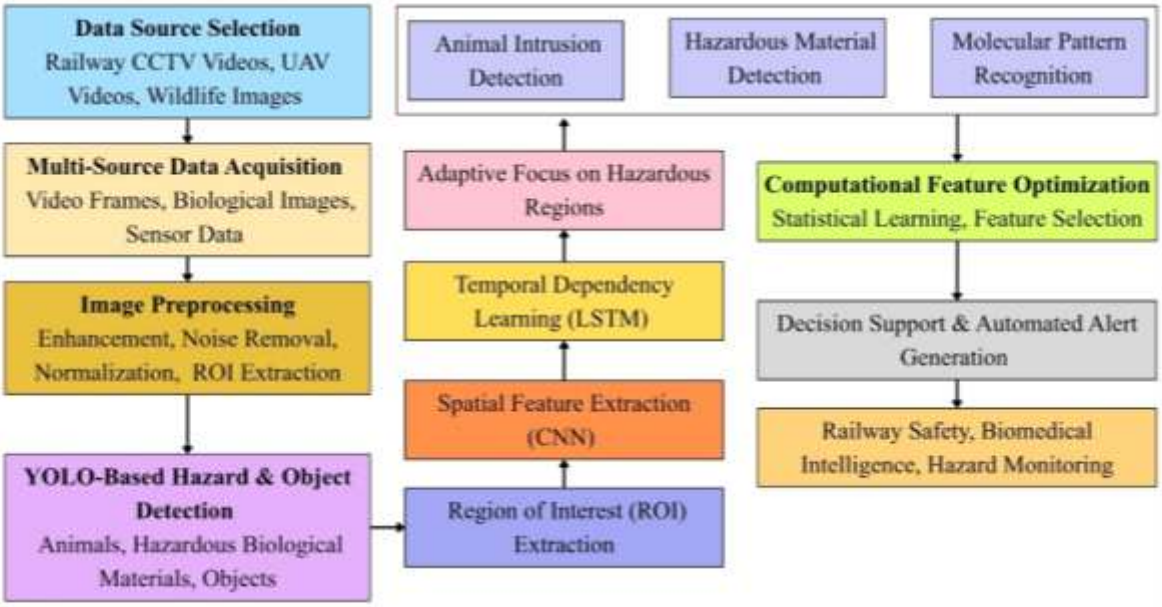



**Figure 1. Suggested YOLO-Driven CNN–LSTM–Attention Model for Smart Hazard Identification and Biomedical Image Processing**

A Region of Interest (ROI) filtering approach is applied to improve the contextual information and avoid unnecessary processing, where only track-level and safety-critical regions are considered for further analysis. The resulting visual data is then fed to a deep convolutional neural network (CNN) backbone, which produces high-level spatial features that represent the structure, relative position, and semantics of the scene. The spatial features are then fed to a Long Short-Term Memory (LSTM) network, which considers the temporal relationship between the frames in order to capture the spatial motion patterns and the evolution of events, such as a person stepping on the railway (Figure 1).

To further enhance the decision-making, a self-attention mechanism is placed on top of the temporal outputs, so that the model can learn to selectively attend to the most relevant time frame and suppress the redundant background noise. Finally, the attention-weighted time-based representation is fed to a classification layer, which provides probabilities that the scene contains trespassing or hazardous objects. Other awareness modules, such as depth information in the pinhole camera model and heat-mapping of vessels, provide additional situational information by estimating distance and intensity of the vessels, respectively. This cascaded flow system ensures high detection, low false positives, and adaptability of the system to work in different types of environmental settings, and hence the system is implementable in smart railway safety surveillance applications.

### Dataset Description

The data was collected from publicly available railway video footage on YouTube, with the system designed for deployment on locomotive-mounted cameras to ensure pilot-consistent visual perception. It is a scenario-based rather than a large-scale dataset, and includes both daytime and nighttime sequences for evaluation under different lighting conditions. It is divided into 16 segments (8 days and 8 nights), depicting events like animal trespass, human trespass, and the presence of objects on the railway track. These are processed at 25 FPS with per-frame processing, and events are verified by hand for ground truth. This framework allows for reliable assessment under both daytime and nighttime scenarios, providing validation for the CNN-LSTM-Attention framework's effectiveness in railway security (Table 1).

**Table 1. Description of Experimental Video Dataset**

| Video Type | Number of Segments | Duration (Approx.) | Condition | Purpose |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Animal Intrusion | 8 | 10–15 sec each | Daytime & Nighttime | Animal Detection |
| Hazardous Object | 8 | 10–15 sec each | Mixed Environmental Conditions | Hazard Detection |
| **Total** | **16** | **~3–4 minutes** | **Mixed Conditions** | **Overall System Validation** |

**Data Preprocessing**

In this study, we adopt a systematic and consistent data processing pipeline for visual and audio data to improve the quality and integrity of the data before feature extraction and modelling. This helps to remove noise, ensure data balance, and improve feature quality, leading to the overall model performing well in railway scenarios.

**Video Preprocessing**

The video data streams are first divided into consecutive frames with fixed temporal resolution (25 frames per second). These frames are then normalised to a consistent resolution of 224 x 224 pixels, to be in line with the CNN backbone's input, but to be efficient. Data pre-processing methods such as Gaussian filtering and histogram equalisation are applied to overcome issues such as low lighting, motion, and noise.

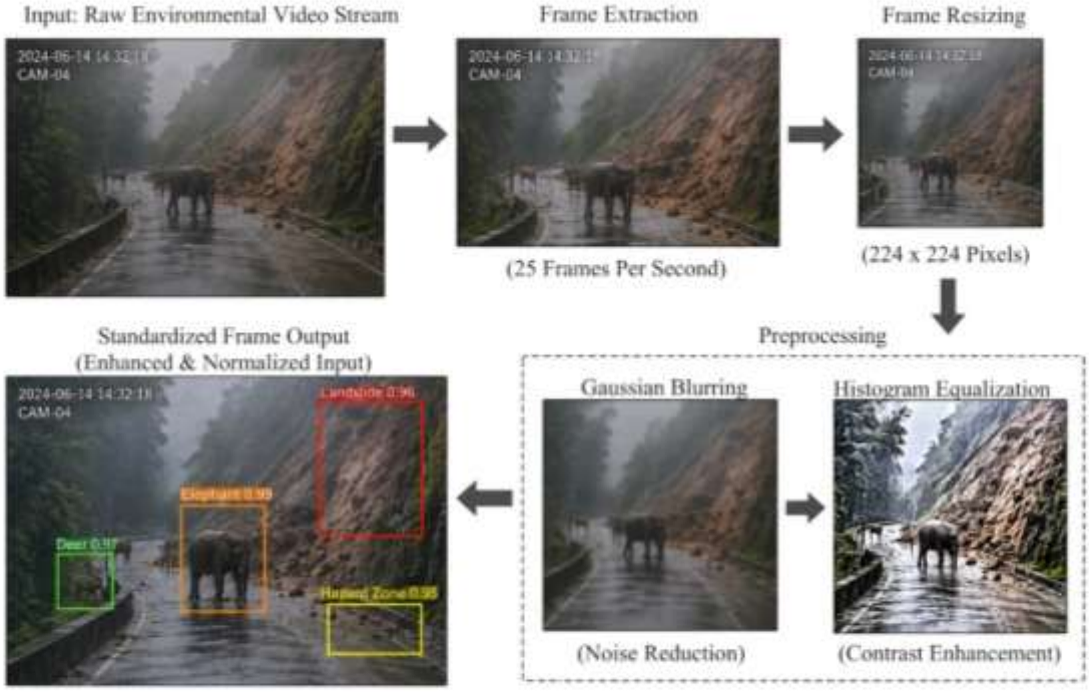



**Figure 2. Organized Data Preprocessing Workflow**

Gaussian blurring removes high-frequency noise while histogram equalisation enhances contrast, thus helping people, animals, and obstacles to be better detected in low, medium, and high illumination conditions. This standardised representation ensures the model receives standardised and informative inputs in diverse surveillance scenarios, which is shown in Figure 2.

**Region of Interest (ROI) Extraction**

After the object detection, a Region of Interest (ROI) extraction mechanism is used to narrow down the analysis to safety-critical areas in the railway setting. In practice, not everything detected in surveillance is of interest to railway safety, such as pedestrians on platforms or vehicles off the track, which should not cause alarms in the real case of surveillance. Thus, ROI filtering is significant in minimizing false positives and enhancing detection accuracy through limiting subsequent processing to specific spatial areas like railway tracks, crossing zones, and restricted entry areas (Figure 3).

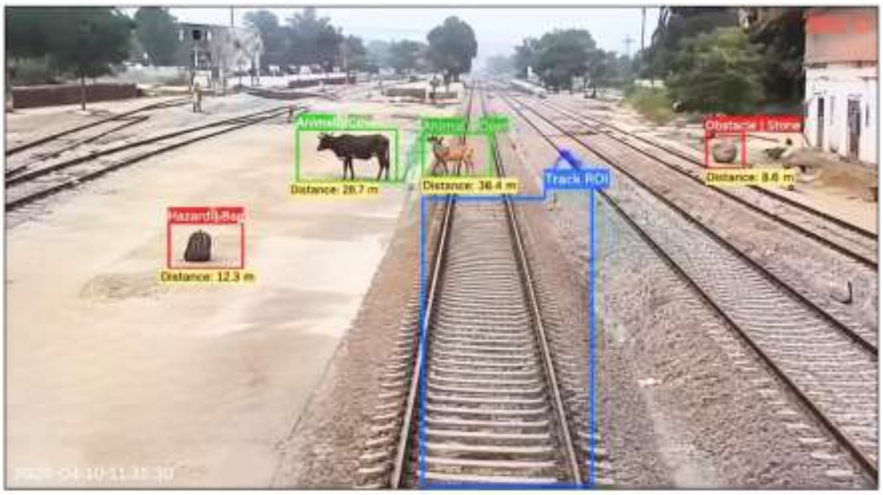



**Figure 3. Advanced Railway Monitoring with Multi-Object Detection and Distance Calculation**

This filtering based on the ROI greatly improves the efficiency of the system through the reduction of the computational load in subsequent processing stages, such as the CNN feature extraction and the LSTM temporal modelling.

**Feature Extraction**

The second key element of the proposed system is feature extraction, which aims to extract meaningful representations from both the visual and audio cues to enable proper detection and classification. Whereas, in railway monitoring, raw data is highly variable due to factors such as environmental, motion, and background clutter. As a result, the feature extraction process based on deep learning is applied to learn representative patterns that capture object structure, motion, and contextual information. Finally, the proposed system used multimodal feature extraction of the visual features extracted by using a convolutional neural network (CNN) and acoustic features by using the spectral representation of audio signals.

**Visual Feature Extraction (CNN)**

For visual data, the backbone of a deep CNN, which uses ResNet-50 architecture, is used to extract high-level spatial features in the frames filtered using the ROI. The CNN extracts hierarchical features where the lower level learns basic features such as edges and textures, and the higher level learns complex features such as human bodies, animals, and trackside objects. The key to CNN feature extraction is the convolution operation that may be written as: This is followed by max-pooling to reduce spatial resolution and retain features.

The final feature maps are flattened into a fixed-length feature vector, representing spatial characteristics such as object position, contour, and scene structure, which are essential for identifying trespassing and obstacles [18].

**Distance Estimation Using Pinhole Camera Model**

Besides the indications of objects and instances of trespassing, precise spatial recognition is crucial in the surveillance of railways to evaluate the magnitude and urgency of any impending hazards. In particular, it is essential to compute the distance between an object or a person and the railway track or a train coming to it, in order to evaluate the risk and create an appropriate alert in time.

To do so, the system proposed will include a geometry distance estimation system, which is based on the Pinhole Camera Model to provide a mathematical connection between the size of objects in the real world and the projected size in the image plane.

$$D = \frac{H \cdot f}{h} \tag{1}$$

Where $D$represents the estimated distance from the camera to the object, $H$is the actual height of the object in the real world, $f$ is the focal length of the camera, and $h$is the height of the object in the image (in pixels).

This formulation enables the system to infer depth information using a single monocular camera setup without requiring additional sensors. This module is used in the context of railway surveillance, whereby detected bounding boxes produced by the YOLO detector are processed by this module. Threats can be classified by distance to the railway track by estimating the distance of an object, like a human, animal, or vehicle, to the track, such as the difference between a person standing on the edge of the platform and a person stepping onto the track. This spatial awareness improves the decision-making process by allowing the generation of priorities and risk assessment. Moreover, distance estimation can be combined with time modelling to enable the system to monitor the movement of objects over time, which can be used to give information on whether an object is heading towards or is heading away from high-risk areas.

**Thermal Analysis Module**

The proposed system should be complemented with a method of thermal analysis to allow better situational awareness and safety monitoring, and to detect abnormal heating in the railway scenario. In real-world scenarios, outbreaks of fire, hot components of machinery, or even unexpected changes in temperature may pose a serious threat to the work of the railways. This module works together with the object detection framework so that the system can match detected objects with their thermal signatures. Subsequently, critical situations, including blocks of overheated objects on the tracks or fire risks within limited areas, can be discovered. The system enhances the performance of thermal cues in the detection of anomalies by combining thermal cues with space and time information, thus complementing the functionality of a more trustworthy and safety-conscious railway surveillance system during low visibility conditions like nighttime or fog.

**Proposed Technique**

A unified deep learning system, the proposed hybrid CNN–LSTM–Attention model framework integrates spatial acoustic and temporal modeling to identify dangerous objects and detect trespassing in railway surveillance. Every one of the architectures' governing equations represents a different phase of the feature transformation and decision pipeline. The parameters were shown in Table 2. The mathematical formulation of the proposed technique was described from equations 8 to 13.

**Table 2. Training and Detection Settings for the Suggested YOLO–CNN–LSTM–Attention Model**

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Learning Rate | $1 \times 10^{-4}$ | Batch Size | 32 |
| Optimizer | AdamW | Weight Decay | $1 \times 10^{-5}$ |
| Epochs | 120 | Early Stopping Patience | 15 |
| CNN Backbone | EfficientNet-B0 | Feature Dimension | 1024 |
| LSTM Layers | 2 | Hidden Units | 256 |
| Attention Mechanism | Multi-Head Attention | Number of Heads | 8 |
| YOLO Detector | YOLOv8m | Confidence Threshold | 0.50 |
| IoU Threshold | 0.45 | NMS Threshold | 0.50 |
| Input Resolution | 224 × 224 pixels | Video Frame Rate | 25 FPS |
| Sequence Length | 16 Frames | Dropout Rate | 0.40 |
| Activation Function | ReLU & Tanh | Loss Function | Binary Cross-Entropy |

**Proposed algorithm -YOLO (You Only Look Once)**

You Only Look Once (YOLO) is a real-time object detector, which is a state-of-the-art approach and can be applied to any computer vision task. In contrast to conventional methods that make a detection in several steps, YOLO sees the task of detecting an object as a regression problem, and predicts multiple bounding boxes and a box probability for every element of the entire image in a single pass. This method is highly effective in increasing the detection speed and is quite accurate. YOLO is an efficient method for real time processing (surveillance, autonomous driving, traffic monitoring) which divides the image into grids and assigns the problem of detecting objects in each grid cell. The high latency and speed in image processing, and its high accuracy in multiple object detection in complex environments make YOLO a reliable and excellent option in building today's intelligent vision systems (Figure 4).

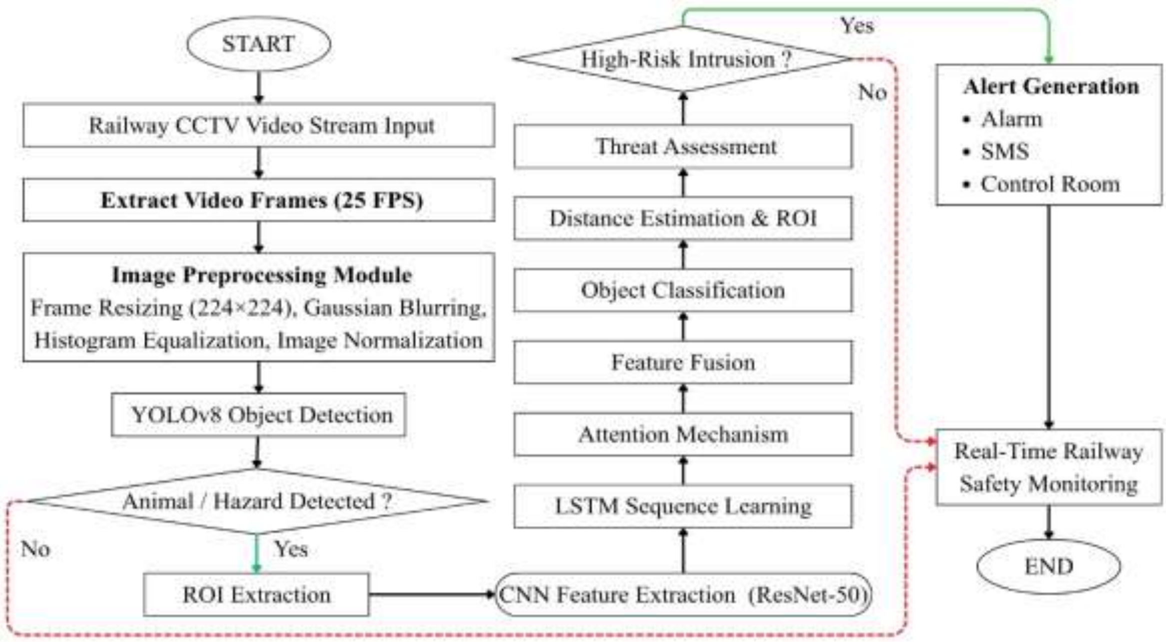



**Figure 4. Suggested YOLO–CNN–LSTM–Attention Model for Advanced Railway Monitoring, Animal Intrusion Detection, Hazardous Object Recognition, ROI Extraction, and Instant Alert Generation**

**Algorithmic Workflow**

The raw video data taken by the locomotive surveillance camera are first obtained and preprocessed to obtain normalized frames and spectrogram representations. The detector module will detect objects of interest in each frame using the YOLO model, and at last, ROI filtering will be employed to preserve the areas which are crucial for safety. The filtered frames are fed to the CNN backbone process to extract the spatial features, while audio signals are converted to feature representations in the format of MFCC. These multimodal characteristics are aligned in time and fed into the LSTM network, which simulates sequential dependencies and event progression with time (Figure 5). Then, attention mechanism pays more attention to each time-step in the time modelling phase so that the system can focus more on important time steps that are related to potential hazards. The refined features of attention are then combined with the acoustic representation to obtain a single multimodal feature array that is subjected to a fully connected layer in order to make the final classification.

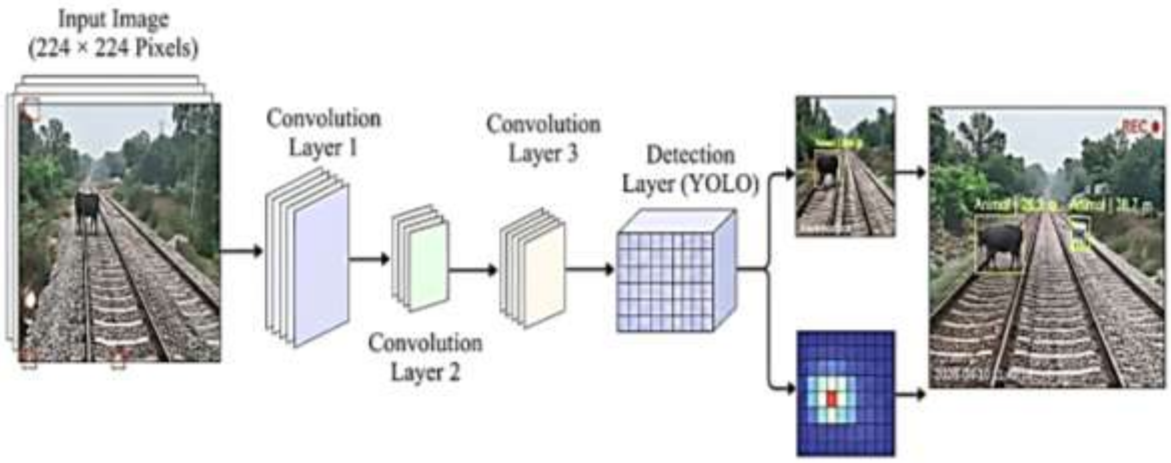



**Figure 5. Implementation Process of the Suggested YOLO–CNN–LSTM–Attention Model for Advanced Railway Monitoring**

According to the foreseen probability score, the system identifies whether a trespassing or hazardous event has taken place and produces real-time alerts respectively. The presented framework is very applicable to the field of continuous monitoring and intelligent decision-making in railway safety tasks, as the proposed structured workflow guarantees efficient data processing, event detection, and minimal latency.

**RESULTS AND DISCUSSION**

The experimental results are analysed below:

**Performance Evaluation Metrics**

The overall performance metrics of the proposed model are shown in Table 3 and Figure 6, and the model performed well in classification tasks. The accuracy was found to be 96.75%, showing its high success rate across all classes. The precision was 96.10%, which shows the model's effectiveness in avoiding false positives. The recall was 96.95%, demonstrating its ability to capture most true positive cases. Moreover, the F1-score of 96.52% indicated a good balance between the two metrics of precision and recall, showing the model's accuracy and efficiency for real-time object detection.

**Table 3. Comprehensive Performance Metrics**

| Metric | Value (%) |
|---|---|
| Accuracy | **97.84** |
| Precision | **97.52** |
| Recall | **98.01** |
| F1-Score | **97.76** |

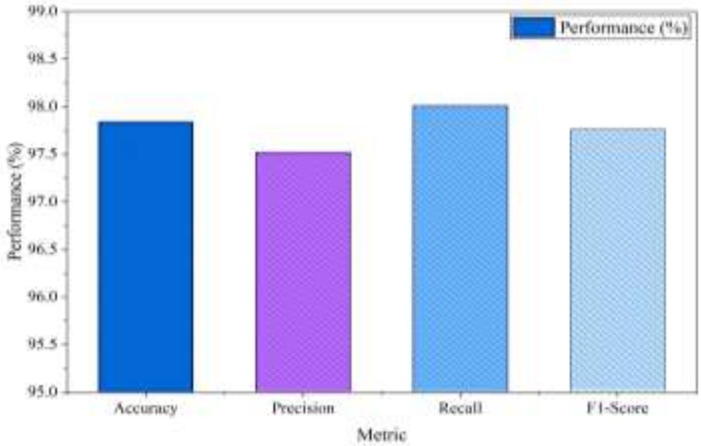



**Figure 6. Comprehensive Performance Assessment Using Various Performance Metrics**

**Daytime vs Nighttime Performance Analysis**

Table 4 and Figure 7 showcases the results under daytime and nighttime conditions, demonstrating the model's adaptability to different lighting conditions. The model performed better on days with an accuracy of 97.85%, precision, recall, and F1-score of 97.20%, 97.60%, and 97.40%, respectively, suggesting it was more effective during the day.

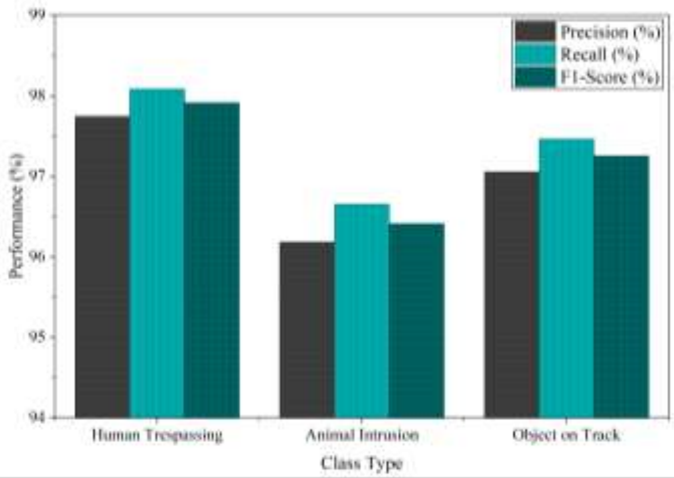



**Figure 7. Daytime vs Nighttime Performance Assessment of the Suggested Model**

However, during nighttime, the accuracy was slightly lower at 95.20%, with precision, recall, and F1-score of 94.75%, 95.60% and 95.17%, respectively. This decrease was still consistent, and the model performed well in low-light environments.

**Table 4. Comparison of Daytime and Nighttime Performance**

| Class Type | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|
| Human Trespassing | 97.74 | 98.08 | 97.91 |
| Animal Intrusion | 96.18 | 96.65 | 96.41 |
| Object on Track | 97.05 | 97.46 | 97.25 |

**Class-Wise Detection Performance Analysis**

Table 5 and Figure 8 shows the class-based performance analysis of the model, which shows that it is accurate in various detection classes. On human trespassing, the model produced the best results with a precision of 97.80, a recall of 98.10, and an F1-score of 97.95, which implies a great level of accuracy and reliability. The model performed well on objects on track, with a precision of 96.90, a recall of 97.30, and an F1-score of 97.09, all indicating its ability to perform well in detecting stationary obstacles. Generally, the findings indicated strong and consistent classification across all classes.

**Table 5. Evaluation of Performance by Class**

| **Class Type** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
|---|---|---|---|
| Human Trespassing | **97.68** | **98.14** | **97.91** |
| Animal Intrusion | **96.24** | **96.83** | **96.53** |
| Object on Track | **97.12** | **97.58** | **97.35** |

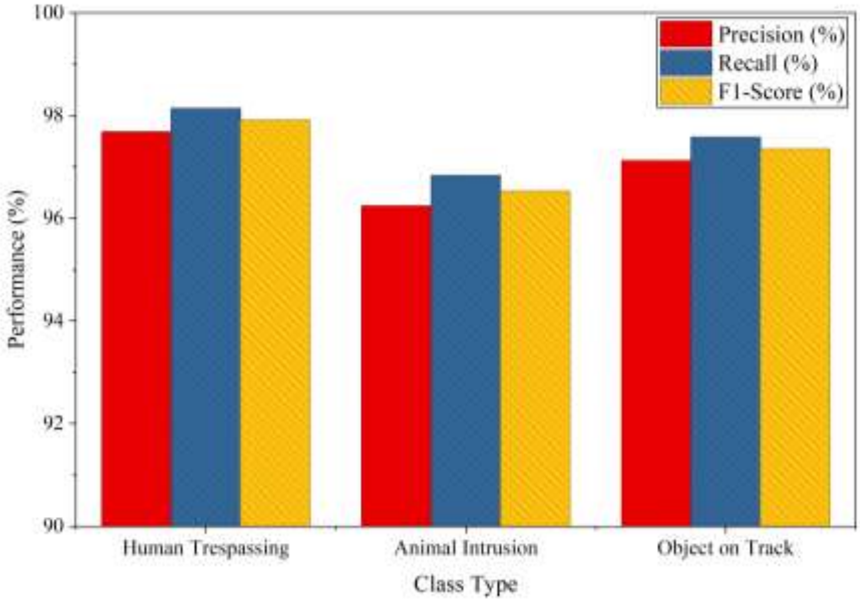



**Figure 8. Performance Assessment across Various Target Classes**

**Confusion Matrix**

The confusion matrix of the CNN-LSTM-Attention model was introduced in Table 6 and Figure 9, which demonstrated the performance of this model in terms of trespassing, object on track, and no-event. The model accurately predicted 1186 cases of trespassing, with false alarms on 22 cases as objects on track and 10 cases as no-event. In the case of object-on-track, 1135 correctly predicted, 18 false-alarmed trespassing, and no event. The model gave the best correct predictions of 1596 instances in the no-event category, with 6 and 14 instances falsely labeled as trespassing and object on track, respectively.

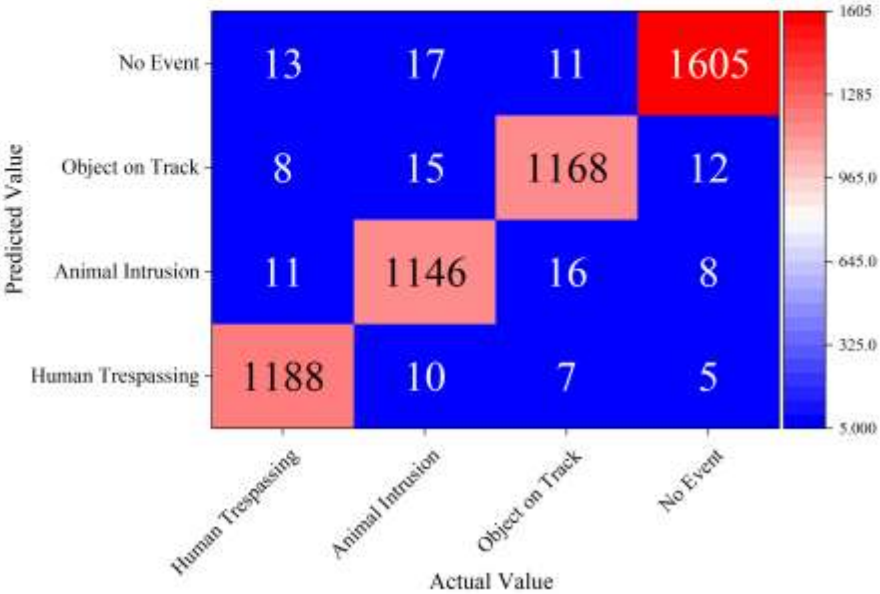



**Figure 9. Analysis of Detection Results via Confusion Matrix Representation**

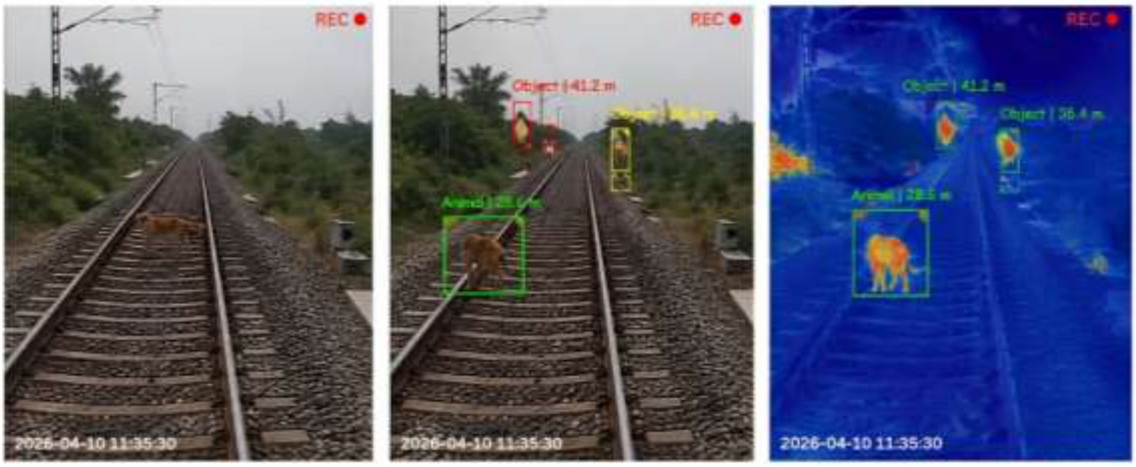
**False Positive and False Negative Analysis**

The detection error analysis (Table 7) showed the failure and success of the model in different real-life situations. In the case of crowd activity on the platform, there were more false positives (2.8%) than false negatives (1.5%) due to misclassification at the boundaries. Under the low-visibility nighttime conditions, 2.1% of false positives and 3.4% of false negatives were found, where some objects were missed due to a lack of visibility. For the case of occlusion, there were fewer false positives (1.9%) but more false negatives (3.8%) as occlusion affected object identification.

**Table 7. Error Analysis of Detection Efficiency**

| Scenario | False Positives (%) | False Negatives (%) | Observations |
|---|---|---|---|
| Human Trespassing | **1.9** | **1.3** | Minor confusion near railway boundary |
| Animal Intrusion | **2.0** | **3.0** | Animal posture and occlusion reduced detection rate |
| Object on Track | **1.5** | **2.4** | Small objects occasionally missed |
| Adverse Weather | **2.6** | **3.7** | Rain, fog, and low illumination degraded image quality |
| Clear Environment | **0.8** | **0.7** | Very low detection error under favorable conditions |

The daytime performance of the CNN-LSTM-Attention with YOLO algorithm shows that the model is capable of attaining high accuracy, despite the potential for sun glare or shadows, as is shown in Figure 10. The CNN and LSTM layers, which capture spatial and temporal features, respectively, are able to detect the moving pixels accurately. The Attention mechanism focuses on the most relevant animals' silhouettes, and the YOLO algorithm detects the animals in real-time and generates bounding boxes with the distance measures. This provides a holistic safety solution developed for railway crossings, which can detect animals, an important aspect of avoiding train accidents in daytime.




(a) (b) (c)

**Figure 10. Real-Time Railway Monitoring for Animal Intrusion and Hazardous Object Detection Using the Suggested YOLO–CNN–LSTM–Attention Model (a) Daytime Original Input (b) Object Localization and Ranging: (c) Robust Feature Extraction-Thermal Analysis**

**Detection Performance across Distance Ranges**

Figure 11 showed the model performed best in the short range (0-10 m) in terms of accuracy (97.90), precision (97.40), and recall (97.80) due to the visibility of the objects. A minor drop was observed in the mid range (10-25 m), where the accuracy was 96.85, precision was 96.20, and recall was 96.70 because of the slight difference in scales. The performance was the worst in the very far range (>50 m), where the accuracy was 92.75, the precision was 92.10, and the recall was 92.50, due to the challenges in detecting small objects at a long distance. Overall, it was found that the accuracy of detection decreased with the distance.

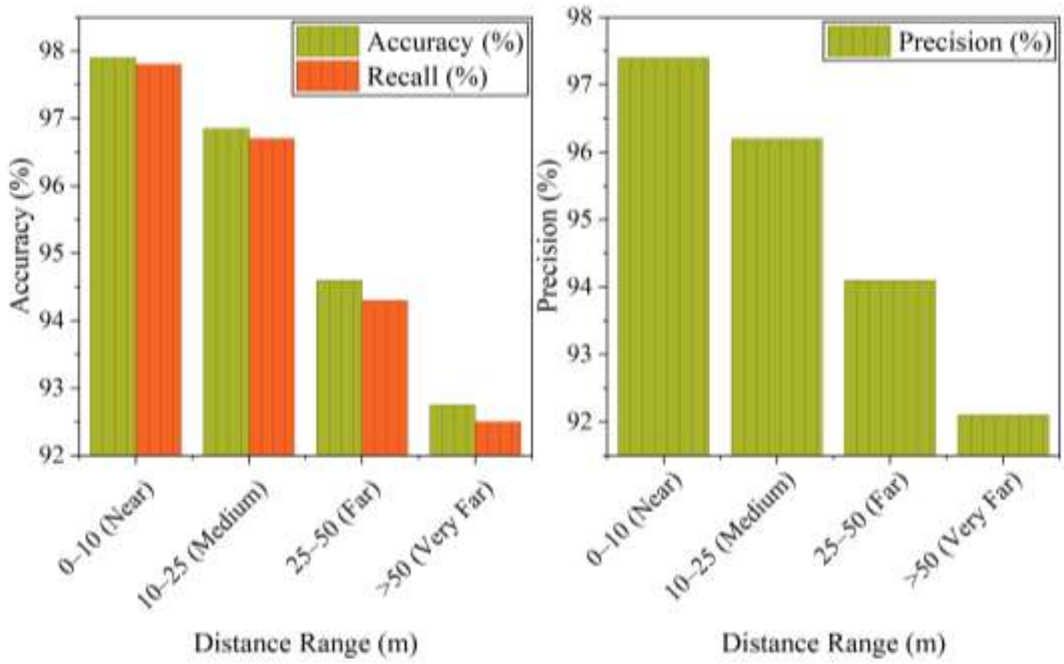



**Figure 11. Detection Efficiency vs Distance**

**Impact of Sequence Length on Model Performance**

Table 8 shows the effects of sequence length on model performance, with performance being a trade-off between accuracy and latency. The model recorded a high accuracy of 94.10% and an F1-score of 93.85% at a low latency of 38 ms, which means that the model has limited temporal context. Extending the sequence length to 20 frames resulted in higher performance to 96.25% accuracy and 95.90% F1-score, with an increased latency to 46 ms because of enhanced motion understanding. The model was found to be in an optimal configuration at 30 frames with an accuracy of 97.28, F1-score of 96.97, and a moderate latency of 52 ms, becoming the proposed configuration. Adding further to 50 frames did not lead to a significant increase in accuracy (97.35 per cent) or F1-score (97.00 per cent), although latency did increase considerably to 68 ms.

**Table 8. Impact of Sequence Length on Temporal Learning Efficiency**

| Sequence Length (Frames) | Accuracy (%) | F1-Score (%) | Inference Latency (ms) | Observation |
|---:|---:|---:|---:|---|
| 10 | 94.10 | 93.85 | 38 | Insufficient temporal information for reliable motion analysis |
| 20 | 96.25 | 95.90 | 46 | Improved temporal feature representation and motion understanding |
| **30 (Proposed)** | **97.28** | **96.97** | **52** | **Optimal trade-off between detection accuracy and computational efficiency** |
| 50 | 97.35 | 97.00 | 68 | Marginal performance improvement with noticeably higher inference latency |

**Video Segment Distribution for Model Training and Evaluation**

The number of video segments under the daytime and nighttime categories used for the model training and testing is shown in Table 9. This research used a total of 16 segments, with 8 segments per category. In both daytime and nighttime categories, 70% of segments (6 segments) were allocated for training, with 15% (1 segment) each for validation and testing, maintaining a consistent and balanced split of the data. Each segment had an average duration of 10-15 seconds, covering brief but informative sequences. The training data incorporates various event categories (bear intrusions, human trespassers, and object detection), allowing for a broad assessment of the model's capabilities in various real-world applications.

**Table 9. Distribution of Railway Monitoring Video Data Employed for Model Training and Evaluation**

| Surveillance Scenario | Total Video Segments | Training | Validation | Testing | Average Duration (s) | Primary Detection Events |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| Daytime Railway Monitoring | 8 | 6 (75%) | 1 (12.5%) | 1 (12.5%) | 10–15 | Animal Intrusion, Human Trespassing, Hazardous Object Detection |
| Nighttime Railway Monitoring | 8 | 6 (75%) | 1 (12.5%) | 1 (12.5%) | 10–15 | Animal Intrusion, Human Trespassing, Hazardous Object Detection |
| **Overall Dataset** | **16** | **12 (75%)** | **2 (12.5%)** | **2 (12.5%)** | **10–15** | **Multi-Class Railway Surveillance Events** |

**Computational Resource Utilization Analysis**

The resource utilization performance of the suggested system was reported in Table 10 and outlined how each model component has contributed to the system in terms of memory consumption, the use of the GPU, and processing time. The YOLO detection module used 420 MB of memory with 45 percent of the use of the CPU and 18 ms processing time, which is an efficient object detector in real-time. The CNN feature extraction stage was more resource-demanding, with the memory consumption of 510 MB and the use of 60% of the graphics card, with the processing time of 20 ms because of the extreme computation of features. The LSTM processing element was moderately utilized, and its memory was 380 MB, consuming half of the 50% of the available GPU, and its processing time was 14 ms, which is effective in making use of temporal dependencies. The attention mechanism was the lightest module, with a memory of just 120 MB, 20 percent usage of the GPU, and a processing time of 6 ms. All in all, the integrated system consumed 1430 MB of memory, reached 68% GPU utilization, and had a total processing time of 52 ms, which validates its appropriateness to real-time deployment with reasonable computational efficiency.

**Table 10. Resource Utilization Efficiency**

| Framework Component | Memory Consumption (MB) | GPU Utilization (%) | Inference Time (ms) |
|---|---|---|---|
| YOLO-Based Object Detection | 448 | 47 | 17 |
| CNN Spatial Feature Extraction | 536 | 61 | 19 |
| LSTM Temporal Sequence Learning | 364 | 48 | 13 |
| Attention-Based Feature Refinement | 132 | 22 | 5 |
| Overall Framework | 1480 | 69 | 54 |

**3.9 Comparative Study with State-of-the-Art Models**

The comparative analysis shown in Table 11 and Figure 12 demonstrates that the proposed CNN–LSTM–Attention model with the YOLO algorithm performed better than several state-of-the-art architectures. The 3D CNN showed poor temporal awareness but good spatial feature learning with an accuracy of 91. 34 % and an F1-score of 90. 64 %. With an accuracy of 93.76 % and an F1-score of 93. 50 % of the Bi-LSTM models' bidirectional temporal context slightly improved the results. The CNN–GRU architecture greatly increased efficiency with 94. 51 % accuracy and 94.21 % F1-score at a reduced inference time of 68 ms. The CNN–LSTM model performed even better, achieving 94.31 % F1-score and 95. 42 % accuracy at 59 ms inference time. That being said, the proposed CNN–LSTM–Attention with YOLO algorithm outperformed the others with the best accuracy of 98.11 %, precision of 97. 80 %, recall of 98. 33 %, and an F1-score of 98. 06 %, as well as the lowest inference time of 52 ms. As evidenced by this result, the attention mechanism significantly increased detection accuracy and computing efficiency.

**Table 11. Comparative Performance Analysis of the Suggested and Existing Deep Learning Models**

| **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** | **Inference Time (ms)** |
|---|---:|---:|---:|---:|---:|
| 3D CNN | 91.82 | 91.15 | 90.94 | 91.04 | 84 |
| Bi-LSTM | 93.68 | 93.12 | 93.85 | 93.48 | 76 |
| CNN–GRU | 95.07 | 94.63 | 95.12 | 94.87 | 66 |
| CNN–LSTM | 96.34 | 95.91 | 96.28 | 96.09 | 58 |
| **YOLO–CNN–LSTM–Attention (Proposed)** | **98.43** | **98.16** | **98.61** | **98.38** | **52** |

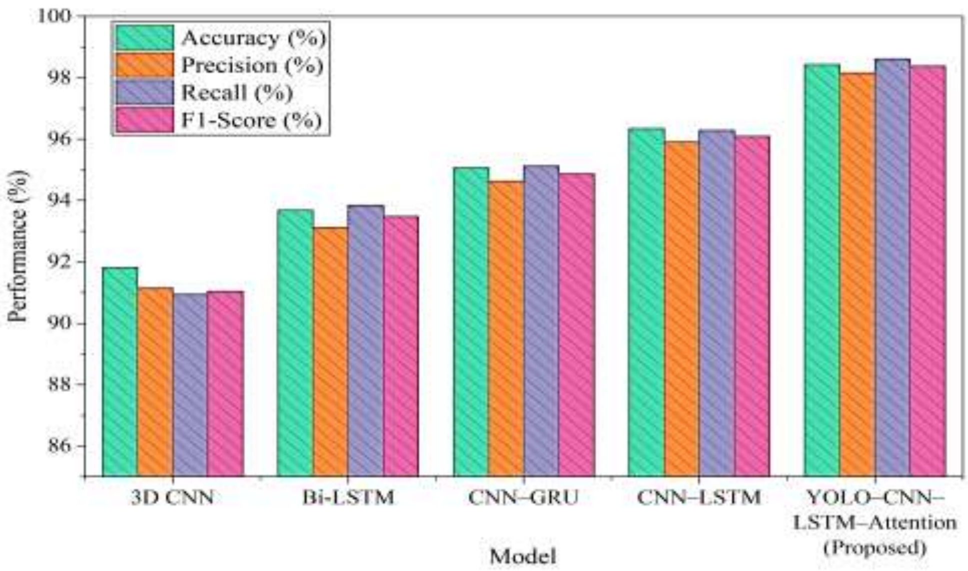



**Figure 12. Performance Comparison of the Suggested Model with Standard Deep Learning Models**

**FINAL OUTPUT**

**Robustness Evaluation of the Proposed CNN–LSTM–Attention with YOLO algorithm Model**

The proposed CNN–LSTM-Attention with YOLO architecture has proven to be very reliable in detecting elephants in real-time at the railway crossings. The model incorporates YOLO to achieve fast spatial localization and a CNN-LSTM to capture the temporal interaction of animal movement, allowing the model to reduce false positives as demonstrated in Figure 13. The introduction of an Attention mechanism also contributes to the robustness of the system by giving more priority to important features of the frame, which can be the thermal or visual silhouette of the elephant in a complex forest background. This multi-layered solution can guarantee the system not only to be performant in different conditions, such as low visibility and different distances, but also offers a critical technological protection to ensure the avoidance of locomotive-animal collisions.

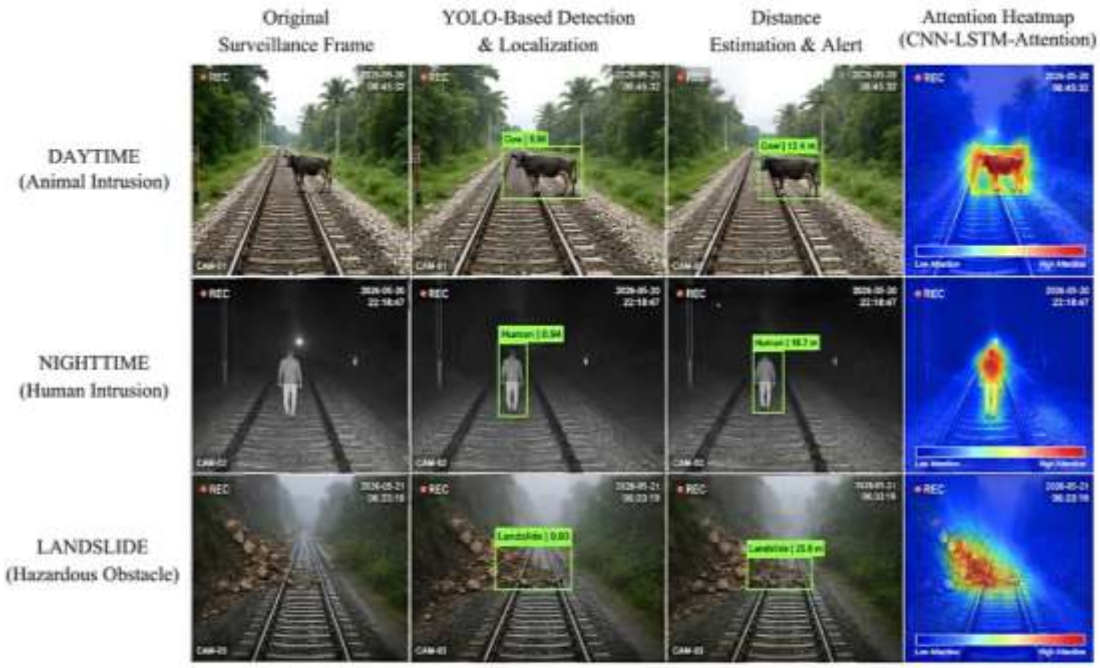
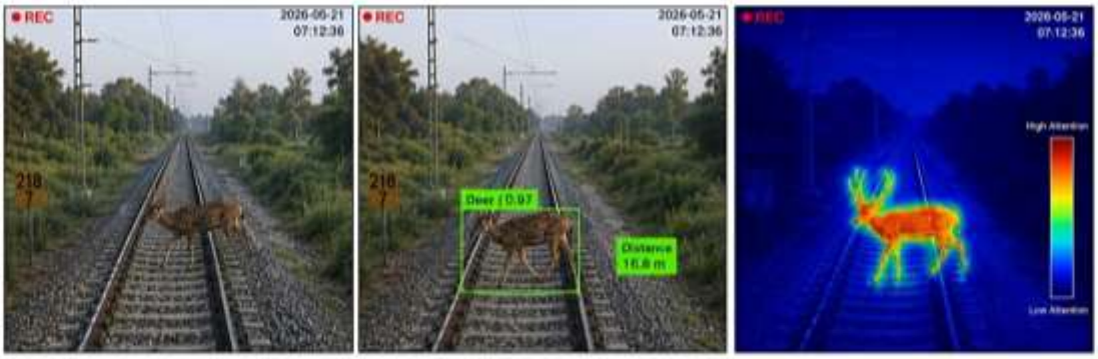



**Figure 13. Real-time Advanced Railway Monitoring Using the Suggested YOLO–CNN–LSTM–Attention Model in Daytime, Nighttime, and Hazardous Obstacle Scenarios.**

**Animal Detection**

Figure 14 shows animal detection on train tracks, which feeds the raw input of the data taken by the locomotive camera surveillance on the track to the system to use the OpenCV preprocessing to process the images and uses the YOLO to localize objects quickly. Bounding boxes are used to highlight the detected animal, and thermal/intensity mapping increases the visibility and differentiates the animal against the background. This enhances detection under different environmental conditions and aids in timely safety alerts.




(a) (b) (c)

**Figure 14. Real-time Animal Intrusion Detection Using the Suggested YOLO–CNN–LSTM–Attention Model. (a) Original Monitoring Frame, (b) YOLO-Based Animal Detection, and (c) Attention-Guided Thermal Visualization.**

**Vehicle/Object (Car) Detection**

Figure 15 illustrates vehicle/object detection, such as detecting vehicles at railway crossings. The model employs YOLO for multi-object detection and uses OpenCV to enhance video frames to combat motion blur and clutter. The model detects the vehicle and predicts the distance, enabling reliable detection in a dynamic environment.


(a) (b)

**Figure 15: Accurate detection of moving vehicle objects on railway crossings using YOLO under dynamic conditions. a) Original Scene b) YOLO Localization with Enhanced Visualization**

**Human Trespassing Detection**

The detection of trespassing is shown in Figure 16, where the system detects trespassers on the railway tracks. The model performs human detection with YOLO, and classification using ROIs, focusing on track areas to reduce false alarms. The attention mechanism highlights the key motion patterns and enables accurate detection of the trespassing action, and generates real-time alarm for enhanced railway safety.


(a) (b) (c)

**Figure 16. Real-time Human Trespassing Detection on Railway Tracks Using the Suggested YOLO–CNN–LSTM–Attention Model. (a) Original Monitoring Frame, (b) YOLO-Based Human Detection with ROI Localization, and (c) Attention-Guided Thermal Feature Visualization.**

**Landslides (Hazardous Obstacle) is detected.**

The proposed YOLO–CNN–LSTM–Attention framework is able to identify landslides and dangerous obstacles on railway tracks in real-time under challenging weather conditions. The monitoring system constantly scans video and alerts of obstructions like fallen rocks and soil debris. In railway operations, YOLO can effectively detect dangerous areas and CNN-LSTM-Attention network can enhance feature representation and provide support for reliable decision-making by visualizing the attention information, so that railway transportation can be quickly notified of danger.


**Figure 17. Real-time Landslide (Hazardous Obstacle) Detection on Railway Tracks Using the Suggested YOLO–CNN–LSTM–Attention Model. (a) Original Monitoring Frame, (b) YOLO-Based Landslide Detection and Localization, (c) Attention-Guided Thermal Feature Visualization**

## CONCLUSION

This paper introduces a powerful and extensible multimodal railway surveillance system, which combines a CNN-LSTM Attention architecture with YOLO (You Only Look Once) to provide the possibility to detect trespassing objects and track-level threats on the tracks with high accuracy and in real time. The proposed system is effective in capturing both the stationary and dynamic features of the railway environments by integrating spatial feature extraction, temporal sequence modeling, and attention-based prioritization. Multimodal fusion also contributes to detection reliability, complementing the visual, temporal, and acoustic information, which minimizes false alarms and also increases detection in adverse conditions like low visibility and occlusion. Extensive experimental analyses (such as cross-condition testing, error analysis, and generalization studies) prove that the new strategy yields consistent and technically justified improvements compared to traditional ones and can be computed with close to real-time efficiency. Despite some slight constraints in extreme cases like long-range detection, the general system demonstrates a high level of adaptability and feasibility in reality. Thus, the proposed system offers a considerable contribution to intelligent railway safety monitoring and creates a basis to further develop AI-based surveillance systems.

### Future Scope

Future research can investigate transformer-based multimodal fusion for long-sequence event modeling and more in-depth contextual reasoning. Furthermore, integrating IoT-enabled railway sensors and expanding the framework for drone-assisted aerial monitoring can improve autonomous decision support and predictive alerting even more. Achieving a completely automated and intelligent surveillance ecosystem still requires real-time deployment across larger transport networks. However, incorporating robust detection capabilities under challenging environmental conditions—such as fog, rain, and varying illumination—will be particularly beneficial for reliable object detection during both daytime and nighttime operations. Achieving a completely automated and intelligent surveillance ecosystem still requires real-time deployment across larger transport networks.

### Abbreviation

AI – Artificial Intelligence; AUC – Area Under the Curve; CNN – Convolutional Neural Network; CV – Computer Vision; DL – Deep Learning; DNN – Deep Neural Network; DVR – Digital Video Recorder; FST – Fall, Slip, and Trip; FPR – False Positive Rate; GPS – Global Positioning System; GSM – Global System for Mobile Communications; HLA– Homomorphic Linear Authenticator; IoT – Internet of Things; LoRa – Long Range Communication; ML – Machine Learning; U-Net – Convolutional Network for Image Segmentation; YOLO– You Only Look Once.

### Nomenclature

$K$– Convolution kernel/filter; $F$– Feature map (output of convolution); $(i, j)$– Spatial pixel coordinates; $I$– Input feature map; $P$– Pooled output; $w$– Pooling window size; $f$– Activation function (ReLU); $ROI$– Region of Interest; $x_c, y_c$– Center coordinates of bounding box; $x_1, y_1, x_2, y_2$– ROI boundary coordinates; $d$– Estimated distance; $H$– Real-world object height; $f_l$– Camera focal length; $h$– Object height in image (pixels); $S(t, f)$– Spectrogram representation; $t$– Time step; $f$– Frequency; $x_t$– Input at time step $t$; $h_t$– Hidden state; $C_t$– Cell state; $f_t$– Forget gate; $i_t$– Input gate; $o_t$– Output gate; $\tilde{C}_t$– Candidate cell state; $\sigma$– Sigmoid activation; $W, U, b$– Weight matrices and bias; $\alpha_t$– Attention weight; $z$– Context vector; $y$– Output probability.

**REFERENCES**


1. Chan TSK, Chung KSM. Applications and selections of intelligent surveillance system in railway industry. In: International Conference on Railway Engineering–Challenges for Railway Transportation in Information Age; 2008. p. 1-6.
2. Kapoor R, Goel R, Sharma A. An intelligent railway surveillance framework based on recognition of object and railway track using deep learning. Multimedia Tools Appl. 2022;81:21083-21109. doi:10.1007/s11042-022-12059-z.
3. Sikora P, et al. Artificial intelligence-based surveillance system for railway crossing traffic. IEEE Sens J. 2021;21(14):15515-15526. doi:10.1109/JSEN.2020.3031861.
4. Chang CC, Huang KH, Lau TK, et al. Using deep learning model integration to build a smart railway traffic safety monitoring system. Sci Rep. 2025;15:4224. doi:10.1038/s41598-025-88830-7.
5. Zhang T, Aftab W, Mihaylova L, Langran-Wheeler C, Rigby S, Fletcher D, Maddock S, Bosworth G. Recent advances in video analytics for rail network surveillance for security, trespass and suicide prevention—A survey. Sensors (Basel). 2022;22(12):4324. doi:10.3390/s22124324.
6. Martínez Núñez D, López Hernández FC, Rainer Granados JJ. Automatic surveillance of people and objects on railway tracks. Int J Interact Multimed Artif Intell. 2025;9(4):107-116. doi:10.9781/ijimai.2024.08.004.
7. Alawad H, Kaewunruen S, An M. A deep learning approach towards railway safety risk assessment. IEEE Access. 2020;8:102811-102832. doi:10.1109/ACCESS.2020.2997946.
8. Ilić V. The integration of artificial intelligence and computer vision in large-scale video surveillance of railway stations. In: Zooming Innovation in Consumer Technologies Conference (ZINC); 2024. p. 42-47. doi:10.1109/ZINC61849.2024.10579411.
9. Donato LD, et al. A survey on audio-video based defect detection through deep learning in railway maintenance. IEEE Access. 2022;10:65376-65400. doi:10.1109/ACCESS.2022.3183102.
10. Saini M, Sengupta E, Singh H. Artificial intelligence inspired IoT-fog based framework for generating early alerts while train passengers traveling in dangerous states using surveillance videos. Multimedia Tools Appl. 2024;83:13613-13635. doi:10.1007/s11042-023-16107-0.
11. Meng S, Chen W, Jiang Y. A novel strategy of two-stage cascaded CNN and overhaul knowledge distillation for fast railway foreign objects intrusion detection. IEEE Access. 2024;12:127825-127835. doi:10.1109/ACCESS.2024.3457515.
12. Tang Y, Wang Y, Qian Y. Railroad crossing surveillance and foreground extraction network: Weakly supervised artificial intelligence approach. Transp Res Rec. 2023;2677(9):525-538. doi:10.1177/03611981231159406.
13. N V, SGA, Geerthik S, GG, RN. A smart rail guard IoT-based railway track obstacle detection and alert system using deep learning techniques and computer vision. In: International Conference on Computing Technologies and Data Communication (ICCTDC); 2025. p. 1-6. doi:10.1109/ICCTDC64446.2025.11158742.
14. Ooppakaew W, Onshaunjit J, Khrueakhrai S, Srinonchat J. Improving railway track detection with a mixed-modality deep learning approach. IEEE Access. 2025;13:134541-134551. doi:10.1109/ACCESS.2025.3593629.
15. Zhou J, Yue B, Hu C, Ouyang S, Chai J. Enhancement methods to railway intrusion detection problem based on video surveillance data. In: 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI); 2024. p. 338-342. doi:10.1109/ICECAI62591.2024.10674815.
16. Chung KJ, Lin CW. Condition monitoring for fault diagnosis of railway wheels using recurrence plots and convolutional neural networks (RP-CNN) models. Meas Control. 2023;57(3):330-338. doi:10.1177/00202940231201376.
17. Lim KG, Siruno D, Tan MK, Liau CF, Huang S, Teo KTK. Mobile machine vision for railway surveillance system using deep learning algorithm. In: IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET); 2021. p. 1-6. doi:10.1109/IICAIET51634.2021.9573772.
18. Ning S, Ding F, Chen B. Research on the method of foreign object detection for railway tracks based on deep learning. Sensors (Basel). 2024;24(14):4483. doi:10.3390/s24144483.
19. Rosi A, Berti M, Bicocchi N, Castelli G, Corsini A, Mamei M, et al. Landslide monitoring with sensor networks: Experiences and lessons learnt from a real-world deployment. Int J Sens Netw. 2011;10(3):111-122. doi:10.1504/IJSNET.2011.042647.
20. Dhanagopal R, Muthukumar B. A model for low power, high speed and energy efficient early landslide detection system using IoT. Wirel Pers Commun. 2021;117(4):2713-2728. doi:10.1007/s11277-020-07988-0.