Smart Railway Safety: Integrating Deep Learning with Vision Transformers for Obstacle Detection and Track Health Monitoring

CHATURYA GANNE, AVINASH BEJJAM,
AKSHAR CHINTALAPALLY,
NAMRATHA REDDY GADDAM,
VIPUL THOTA, PRAFULLA KALAPATAPU
and VENKATA DILIP KUMAR PASUPULETI

ABSTRACT

Railway safety is critical for preventing accidents caused by obstacles present on the track and infrastructure failures. Unlike, other monitoring systems that often solely focus on either obstacle detection or track condition analysis, limiting their ability to provide a comprehensive risk assessment. This research paper presents a multi-modal deep learning framework where a real-time obstacle detection was performed with the use of the YOLO model. However, detecting an obstacle alone is not sufficient for assessing the collision risk; therefore, a Vision Transformer was incorporated to further refine the detection. Vision Transformers enable a detailed understanding of the obstacle type and characteristics; thereby assisting in increasing the efficiency of impact evaluation. Simultaneously, the condition of the tracks was analyzed using Mask R-CNN alongside YOLO to detect cracks on sleepers and identify loose or broken fasteners. A custom dataset consisting of one-minute track videos, where each frame was converted into images, was used for model evaluation. This multi-modal approach ensures high precision in identifying obstacles as well as the track condition. To evaluate the outputs from obstacle analysis and track monitoring, a fusion model is used to aggregate the outputs, resulting in a risk-weighted decision score. This score assists in deciding the next possible action for the train to take while operating. The proposed system shows higher accuracy, improves the railway monitoring and minimises accident risks, making it suitable for real-time deployment in railway networks.

INTRODUCTION

In many countries, railways are an important mode of transportation. Since their establishment in 1825 [1], they have played a pivotal role in facilitating urban and intercity travel, driving economic progress, and connecting communities across expansive regions, especially in India. India's extensive railway network supports millions of commuters and enables large-scale goods transport across diverse terrains. As modern railways adopt higher speeds and greater capacity, ensuring safety has become increasingly critical. However, these advancements introduce new safety challenges that demand smarter solutions.

Chaturya Ganne[1], Avinash Bejjam[1], Akshar Chintalapally[1], Namratha Reddy Gaddam[1], Vipul Thota[1], Prafulla Kalapatapu[2], Venkata Dilip Kumar Pasupuleti[2]
[1]UG Student, [2]Faculty
Mahindra University, Hyderabad, India
se21ucse044@mahindrauniversity.edu.in,
se21ucse031@mahindrauniversity.edu.in,
se21ucse012@mahindrauniversity.edu.in,
se21ucse135@mahindrauniversity.edu.in,
se21ucse259@mahindrauniversity.edu.in,
prafulla.kalapatapu@mahindrauniversity.edu.in,
venkata.pasupuleti@mahindrauniversity.edu.in

One such incident illustrating the consequences of an inadequate railway safety systems occurred on December 21, 2020, around 2 am, near Bhabanipali village, in Odisha, when a Puri-Surat train collided with an elephant crossing the tracks at night, leading to the elephant's death and the derailment of several wagons. Despite prior caution issued by railway authorities regarding elephant movements and the installation of signboards marking elephant crossing zones, the lack of real-time, on-spot detection has resulted in this causality [2].

Such incidents highlight the severity of accidents caused by wildlife crossings, foreign object intrusions, and structural defects in railway tracks. Consequently, ensuring the operational efficiency and automation of modern railways also requires a shift from conventional, labour-intensive inspections to automated, real-time monitoring systems. Hence, the oversight of critical conditions has driven the adoption of vision-based deep learning frameworks reaffirming the necessity for object detection systems capable of providing immediate alerts to prevent such tragedies.

One such framework is the YOLO (You Only Look Once) architecture, which has gained widespread recognition due to its speed and accuracy. Liu et al. [3] demonstrated the capability of YOLOv3 in monitoring thermite welded joints, showcasing its effectiveness for predictive maintenance. Furthermore, Chang et al. [4] successfully integrated YOLOv3 with Mask R-CNN, leveraging the strengths of both models to enhance intrusion detection and accurately delineate track boundaries. These examples reflect the growing impact of computer vision in addressing diverse railway safety challenges.

Despite these advances, many current implementations treat obstacle detection and infrastructure assessment as separate tasks, leading to incomplete situational awareness. To bridge this gap, a unified, multi-modal framework is proposed, capable of simultaneously analyzing track conditions and identifying potential threats along the railway path. This approach is reinforced by emerging architectures that combine visual and textual reasoning, such as Vision Transformers (ViTs). A notable application in this domain is RailTrack-DaViT, which employs a Dual Attention Vision Transformer to classify rail faults with greater precision than traditional convolutional neural networks [5,6].

Expanding on these foundations, the proposed system introduces a comprehensive deep learning architecture that integrates both structural evaluation and obstacle recognition. Initially, the YOLO model is employed to perform rapid object detection on the railway scene, identifying potential obstacles. The detected regions are then passed to the Vision-and-Language Transformer (ViLT), which analyzes both visual features and semantic context to assess the risk associated with each obstacle. In parallel, a hybrid detection module combining YOLOv8 and Mask R-CNN is utilized to identify track-level defects such as cracked sleepers and missing fasteners. Finally, outputs from both branches are synthesized through a fusion model that consolidates information into a unified risk score. This risk score enables real-time decision-making regarding train operations, thereby significantly improving the safety, reliability, and situational awareness of modern railway networks.

BACKGROUND AND RELATED WORK

Traditional railway safety systems have predominantly relied on rule-based approaches that follow predefined standards, procedures, and technical barriers to address known events. While these methods are effective in handling predictable scenarios, they often lack the flexibility needed to respond to unforeseen circumstances [7]. For example, sensor-driven technologies offer real-time insights into track conditions and train operations.

A notable advancement in this domain is Siemens' Trainguard MT, a communication-enabled train control system that operates with wireless communication between the track and the train, enabling continuous train positioning and integrity monitoring [8]. However, such systems often fall short in incorporating advanced data analysis capabilities, thereby limiting their effectiveness in predictive maintenance and real-time decision-making.

Recent studies are highlighting the growing threat posed by foreign object intrusions onto tracks, which can result in serious accidents. To prevent these accidents, traditional methods that involve manual labour and hardware-driven sensors are used. These methods are limited by human error and inability of human to react in real time. In response, deep learning techniques have emerged as promising alternatives, capable of analyzing large-scale sensor and visual data to to enable proactive risk detection [9]. For instance, deep learning and multi-sensor fusion models have been employed to detect foreign object intrusions, enhancing performance across diverse environmental conditions [10]. Similarly, smart surveillance-based railway safety systems using computer vision and image processing have been proposed for real-time detection of anomalies such as wild animals or human intrusions [11]. In the domain of railway track component monitoring, object detection models employing deep learning such as YOLOv3, YOLOv5, and Mask R-CNN have demonstrated outstanding results in real-time railway fault detection [12]. These models have been effectively used to identify missing fasteners, cracked sleepers, defects in thermite weld joints, and other obstructions. Furthermore, YOLOv5, enhanced with FasterNet and attention models, is considered as a good example which has improved detection precision and inference speed [13].

While object detection models excel at visual perception tasks, recent advancements in scene understanding and multi-modal learning have been driven by the introduction of Vision Transformers (ViTs) and Vision-Language Models (VLMs). Unlike traditional convolutional neural networks CNNs, ViTs capture both global and local contextual features by modelling long-range dependencies within images. One prominent application is ViLT (Vision-and-Language Transformer), which processes both image and text inputs with equal emphasis, allowing for a more comprehensive interpretation of complex scenarios.

Building upon this foundation, VinVL (Visual-Language model) enriches visual representations with strong backbones, substantially boosting performance in vision-language tasks [14]. Similarly, the OSCAR framework effectively maps railway track images with corresponding textual reports, a feature valuable for tasks requiring high level reasoning, such as interpreting maintenance reports alongside visual evidence [15]. Given these developments, the proposed paper adds onto the advancements by presenting a model with dual capability not only reduces the risks

associated with train collisions and derailments but also supports the shift toward smarter, autonomous railway systems.

METHODOLOGY

Data Collection and Preprocessing

The data collection process for this study was designed to support a multi-modal fusion system involving YOLOv8, ViLT, and Mask R-CNN. The initial phase involved utilizing a pre-trained YOLOv8 model, trained on the COCO dataset, to perform object detection across various classes. To train the ViLT (Vision-and-Language Transformer) model, a curated dataset comprising 4000 images from the Animals10 dataset was assembled [16]. This dataset included animal classes cat, dog, cow, sheep, elephant, lion, tiger, and horse with 500 images per class. In addition to animal images, the dataset was also included with images of people, unobstructed railway tracks, and various vehicle types, which were sourced from Roboflow. This led to a total dataset size of approximately 5500 images, which was split into 80% training and 20% validation sets.

All images were resized to 224 × 224 pixels to comply with the input specifications of the ViLT model and to facilitate efficient training. Model generalization capabilities were further evaluated using a separate test set comprising unseen images across all relevant categories. One significant challenge addressed during preprocessing was the scarcity of unobstructed railway track images, with only 40 available initially. To overcome this limitation, the Albumentations library was employed for data augmentation, ultimately increasing the unobstructed track image count to 480.

Furthermore, to facilitate track health assessment, the system leveraged two additional datasets: a YOLOv8-compatible dataset comprising 1000 annotated images of various railway track components, entirely collected through field visits and on-site image capturing; and a Mask R-CNN-compatible dataset consisting of 1700 annotated images of sleeper cracks and structural defects, created using a combination of field-collected data and publicly available annotated images from Roboflow. This diverse, well-structured dataset collection ensures that each model receives task-specific, pre-processed input, thereby helping the decision making in the fusion model.

Proposed Workflow of the Model

To improve overall track health assessment in railway systems, a multi-modal fusion framework is introduced, integrating object detection, visual reasoning, and defect segmentation models, achieving a balanced trade-off between accuracy, interpretability, and real-time response. The overall architecture of the proposed pipeline is illustrated in Figure 1. The framework consists of four major components. First, a pretrained YOLOv8 model is used for real-time obstacle detection, leveraging its efficiency and accuracy in object localization [17]. Second, a fine-tuned Vision-and-Language Transformer (ViLT) is incorporated to understand the surrounding scene; this model processes both image patches and textual queries using transformer layers without relying on convolutional backbones [18]. Third, a YOLOv8 combined

with Mask R-CNN pipeline is employed for detailed track condition analysis, including the identification and segmentation of fasteners, missing fasteners, sleepers, and the rail track. Finally, a Fusion model integrates outputs from the above components to perform risk assessment and aid in informed decision-making. The integration of heterogeneous modalities follows principle established in multimod al machine learning research, which highlights the importance of effective feature representation and cross modal information alignment [19,20].

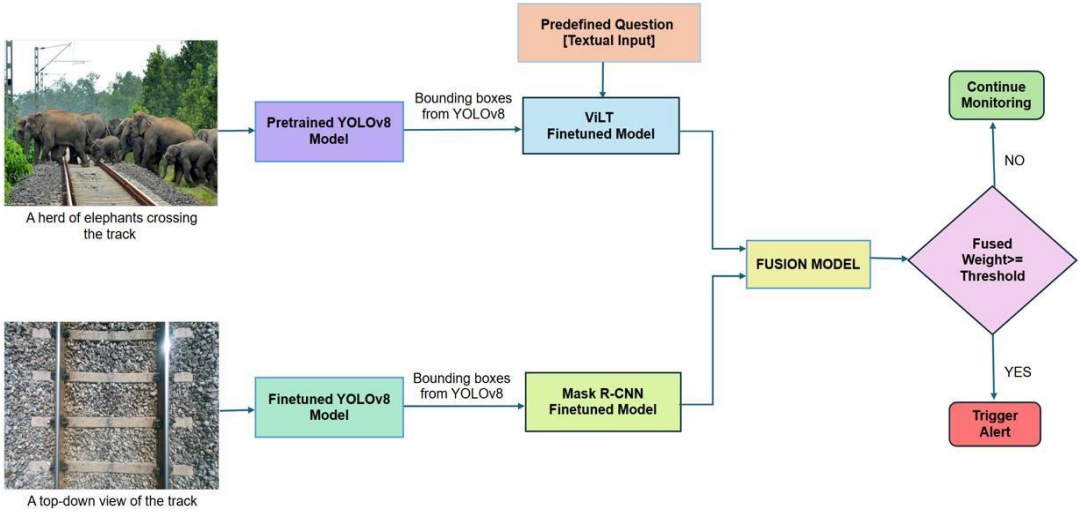


Figure 1: Proposed Architecture

The process begins with capturing an image containing a potential obstacle on the railway track. This image is processed by the pretrained YOLOv8 model, which detects obstacles and generates bounding boxes along with associated confidence scores. These detected regions, together with relevant textual queries, are passed to the ViLT model for vision-and-language reasoning. ViLT performs Visual Question Answering (VQA) to evaluate the potential consequences of a collision with the identified object. It outputs a semantic response that is assigned a confidence score indicating the severity of risk, which is subsequently used in the Fusion module for decision-making [21].

Concurrently, a top-down view of the railway track is analyzed using the fine-tuned YOLOv8 model to identify and generate bounding boxes for fasteners, missing fasteners, sleepers, and the rail track. The bounding boxes corresponding to fasteners and sleepers are passed to a custom-trained Mask R-CNN model, which overlays segmentation masks to identify defective regions, such as cracks in sleepers or missing fasteners [22].

The outputs from both the obstacle detection and defect segmentation branches are then integrated in a Fusion model. Initially, a rule-based fusion strategy is adopted, where the scores from ViLT, YOLOv8, and Mask R-CNN are combined using a weighted sum, modulated by a tunable hyperparameter $\alpha \in [0, 1]$, set to 0.5 during implementation. This strategy combines the ViLT score with either the YOLOv8 score (for detecting missing fasteners), the Mask R-CNN score (for identifying sleeper cracks), or both in more complex cases.

To further enhance decision-making, especially in ambiguous or complex scenarios, the framework employs a learning-based Neural Network (NN) fusion

model as the final stage. This model takes the ViLT response score and the outputs from YOLOv8 and Mask R-CNN as inputs and passes them through fully connected layers with ReLU and sigmoid activations to produce a normalized risk score. If this score exceeds a predefined threshold, the system raises an alert, signaling that it may not be safe for a train to proceed. Otherwise, normal operation is maintained. This adaptive, data-driven fusion strategy ensures robust, interpretable, and context-aware railway safety monitoring under dynamic operational conditions.

Evaluation Metrics

To evaluate the ViLT model, a set of classification metrics, including precision, recall, and F1-scores were computed from a confusion matrix, which summarizes both correct and incorrect predictions, highlighting instances of misclassifications. Precision measures the proportion of correctly predicted instances among all predictions for a given class, while recall quantifies the model's ability to identify all relevant instances. The F1-score, calculated as the harmonic mean of precision and recall, balances both false positives and false negatives. In the evaluation, the ViLT model achieved an F1-score of 1.0 across all classes, indicating minimal misclassification, alongside a validation accuracy of 99.74%. Both macro and weighted averages of these metrics further confirmed robust performance across classes. The high macro averages demonstrate consistent performance even on classes with fewer samples, while the high weighted averages reflect strong performance on classes that appear more frequently. These results indicate that the model does not disproportionately favor dominant classes and generalizes effectively across diverse risk categories and obstacle types in railway environments.

Similarly, the Fusion model, trained using 2000 samples of ViLT outputs combined with YOLOv8 and Mask R-CNN scores over 100 epochs was evaluated. As shown in Figure 2, the fusion model performed well overall, though a few false positives were observed**.** Nevertheless, majority of the predictions were accurate, with the model effectively classifying labels. Future work could focus on training with more extreme or rare cases, which may improve the model's ability to handle edge cases and further strengthen its generalization capabilities.

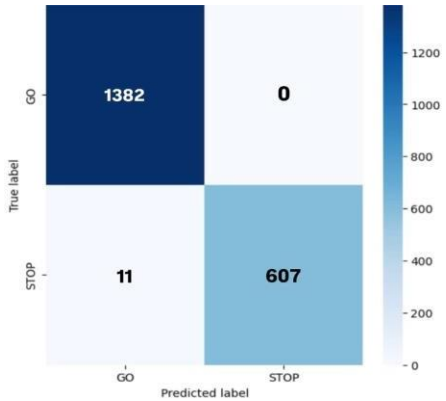
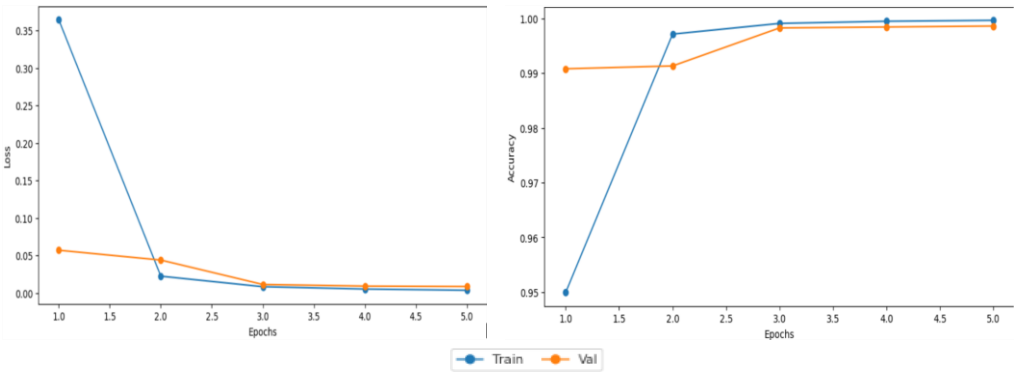


Figure 2. Confusion matrix of the NN-Based Fusion model

RESULTS

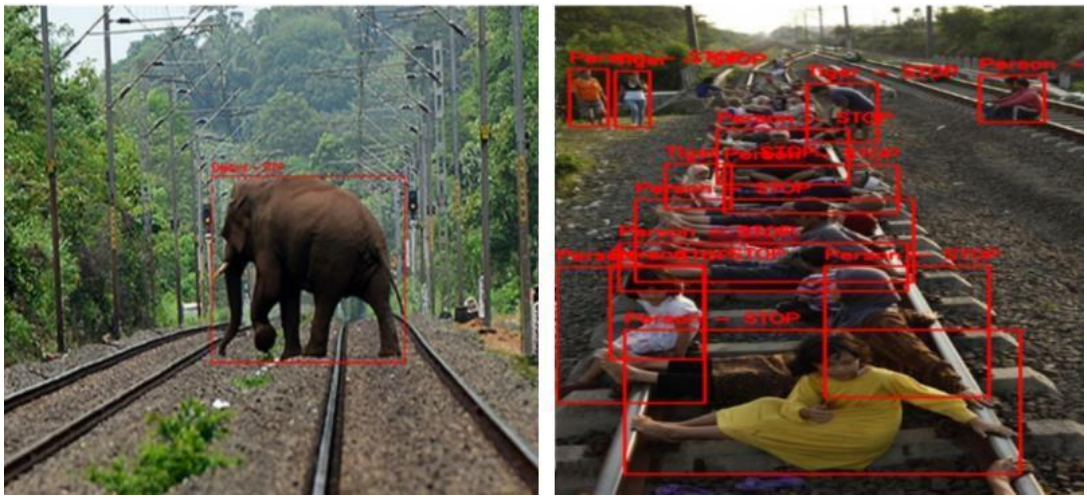
The proposed fusion model, combining ViLT, YOLOv8, and Mask R-CNN, yielded promising results across all evaluated tasks. The ViLT model achieved a validation accuracy of 99.74%, demonstrating its capability to effectively learn and align visual and textual modalities in railway scenarios. The training curves of the ViLT model (Figures 3(a) and 3(b)) have shown a consistent decrease in loss and stabilization of accuracy, indicating proper convergence and effective learning of multimodal representations.



(a) (b)

Figure 3: (a) ViLT Training Loss Over Epochs, (b) ViLT Training Accuracy Over Epochs

Figures 4(a) and 4(b) illustrates various real-world scenarios where the ViLT model accurately classifies an obstacle present on the track.

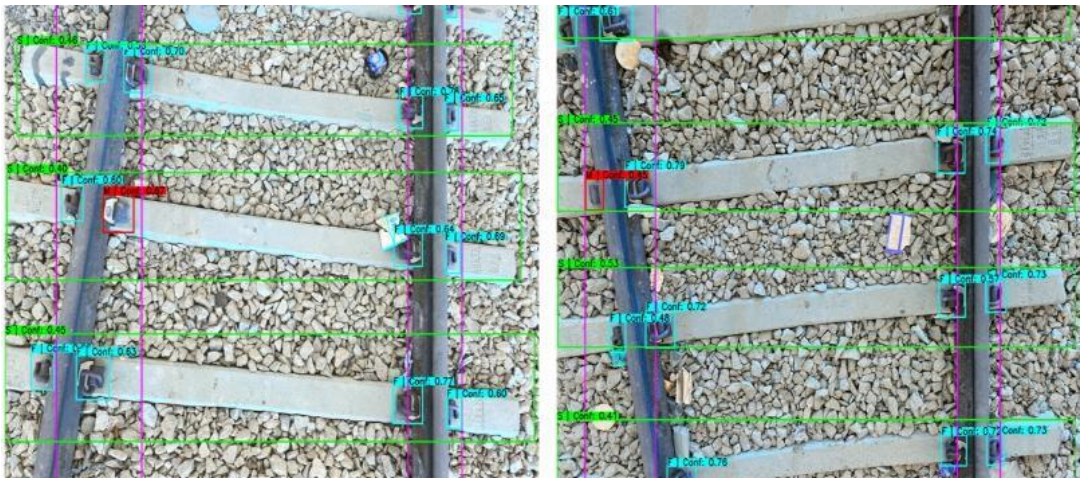


(a) (b)

Figure 4: (a) Inference of the Yolov8 and ViLT Model detected elephant, (b) Detected people

Figures 5(a) and 5(b) showcases top-down view of track components where YOLOv8 detects the missing fasteners, and Mask R-CNN checks for cracks on

sleepers. Based on the fusion logic, the neural network-based fusion model triggers an alert when the fused risk score exceeds the predefined threshold.



(a) (b)

Figure 5. Inference of the Yolov8+Mask R-CNN Model

Through empirical evaluation, the performance of the NN-based fusion model was compared with that of rule-based fusion model. As shown in Table I, the NN-based Fusion model demonstrated high accuracy of 99.45% and precision score of 1.000. However, it displayed a lower recall of 0.9822, and a comparatively high F1 score of 0.99102. These results suggest that the model occasionally struggled to generalize in cases of conflicting predictions, specifically when the ViLT model assigned a high-risk score while the YOLO+Mask R-CNN composite model indicated no risk, or vice versa, leading to a few false positives.

In contrast, the rule-based fusion model attained an accuracy of 83.10% and a high recall of 1.000, successfully detecting all potential hazards, particularly in edge cases, due to its deterministic approach to computing the risk score. However, its precision was lower at 0.6464, indicating a higher occurrence of false positives. Consequently, its F1 score was also lower at 0.7852 when compared to that of the NN-based model.

Therefore, the NN-based model was ultimately chosen as the final solution for the pipeline. In the next phase, the emphasis will be on enhancing the model's robustness and generalization capability by expanding the dataset to include a broader spectrum of edge cases and more complex scenarios.

TABLE I. FUSION MODEL COMPARISION

Fusion Models	Accuracy	Precision	Recall	F1 Score
NN-Based	0.9945	1.0000	0.9822	0.99102
Rule-Based	0.8310	0.6464	1.0000	0.78521

CONCLUSION

This study introduces a multimodal fusion framework for railway track monitoring that integrates ViLT, YOLOv8, and Mask R-CNN models, combining semantic understanding with object detection and segmentation. This approach enables a better interpretation of railway environments by linking contextual reasoning with precise localization of obstacles and structural issues. Evaluation on a custom track dataset demonstrated the effectiveness of the framework, with the ViLT model achieving a validation accuracy of 99.74%, enabling reliable decision-making in critical scenarios.

Extensive testing revealed that the NN fusion model consistently outperformed the rule-based approach, particularly in complex and ambiguous situations where isolated models often struggle. By learning intricate patterns and relationships between modalities, the fusion model demonstrated its adaptability to various track conditions and obstacles. The practical significance of this work lies in its contribution to more intelligent, autonomous railway monitoring systems. By accurately identifying and assessing risks in real-time, the proposed framework supports early detection and mitigation of potential hazards which reduces the likelihood of accidents. While the proposed framework demonstrates promising accuracy and precision, these results are based on a moderately sized, curated dataset. As such, the high performance may not fully reflect behaviour under diverse, real-world conditions. Further testing on larger and more varied datasets is necessary to validate generalizability and ensure robustness in dynamic railway environments.

Future research directions include expanding the dataset to cover a wider range of track anomalies and environmental conditions to further improve the model. Integrating temporal information, such as object movement patterns and track condition evolution over time, will be critical for enabling predictive insights. These advancements are expected to contribute significantly to the development of next-generation Structural Health Monitoring (SHM) systems and more resilient railway infrastructures.

REFERENCES

1. Shaw-Taylor, L., & You, X. (n.d.). The development of the railway network in Britain 1825-1911. Cambridge Group for the History of Population and Social Structure, University of Cambridge. https://www.campop.geog.cam.ac.uk/research/projects/transport/onlineatlas/railways.pdf
2. Mohanty, D. (2020, Dec 21). Elephant dies after being hit by train in Odisha's Sambalpur. Hindustan Times. https://www.hindustantimes.com/india-news/elephant-dies-after-being-hit-by-train-in-odisha-s-sambalpur/story-618lVbkBHjpN6Va2nYuPHO.html
3. Liu, Y., Sun, X., & Pang, J. H. L. (2020). A YOLOv3 -based deep learning application research for condition monitoring of rail thermite welded joints. In Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing (IVSP '20) (pp. 33–38). Association for Computing Machinery. https://doi.org/10.1145/3388818.3388827
4. Chang, C. C., Huang, K. H., Lau, T. K., et al. (2025). Using deep learning model integration to build a smart railway traffic safety monitoring system. Scientific Reports, 15, 4224. https://doi.org/10.1038/s41598-025-88830-7
5. Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. arXiv preprint arXiv:2102.03334. https://arxiv.org/pdf/2004.00849

6. Phaphuangwittayakul, A., Harnpornchai, N., Ying, F., & Zhang, J. (2024). RailTrack -Da ViT: A vision transformer-based approach for automated railway track defect detection. Journal of Imaging, 10(8), 192. https://doi.org/10.3390/jimaging10080192
7. Deboral, C. C., Dhanush, D., Dhyanesh Kumar, S., Prithiv Prakash, A., & Dhivya Lakshumi, S. (2024). Proactive AI-powered railway safety system. International Journal of Creative Research Thoughts, 12(8), 125–135.
8. Khade, S. B., Jadhav, N. V., Chaudhari, M. R., Machpalle, D. R., & Nandargi, K. N. (2024). Rail Safe-Tech: An automatic safety system. International Journal for Research in Applied Science & Engineering Technology, 12(6).
9. Kalinowski, M., & Weichbroth, P. (2023). The sensors-based artificial intelligence Train Control and Monitoring System (TCMS) for managing the railway transport fleet. Rail Vehicles/Pojazdy Szynowe. https://doi.org/10.53502/RAIL-159639
10. Liu, Y., Zheng, Y., Wang, Z., & Zhang, Y. (2020). Research on foreign object intrusion detection for railway safety. In 2020 39th Chinese Control Conference (CCC) (pp. 7061–7066). IEEE.
11. Kumar, H. S. R., Saravanan, T. P., Uz Zaman, A. Z., & Vijayalakshmi, T. S. (2024). Smart surveillance system for railway track safety using image processing. Journal of Real-Time Ima ge Processing, 21(1), 223–235.
12. Huang, A., Yu, H., et al. (2020). Railway safety monitoring based on deep learning object detection algorithm. Scientific Programming, 2020, Article ID 123456.
13. Hu, M., Zhang, Y., Liu, H., & Liang, C. (2024). Enhancing YOLOv5 with FasterNet and attention mechanisms for railway and airway foreign object detection. Multimedia Tools and Applications, 83, 5211–5233.
14. Zhang, P., Li, X., Hu, X., Wang, L., Zhang, Y., Yang, L., & Gao, J. (2021). VinVL: Making visual representations matter in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5579–5588). https://openaccess.thecvf.com/content/CVPR2021/papers/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.pdf
15. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Wang, Y., & Zhou, M. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. arXiv preprint arXiv:2004.06165. https://arxiv.org/pdf/2004.06165.pdf
16. Corrado, A. (2019). Animals-10: A dataset of 10 animal classes. Kaggle. https://www.kaggle.com/datasets/alessiocorrado99/animals10
17. Jocher, G., et al. (2023). YOLOv8: Next-generation real-time object detection. Ultralytics Documentation.
18. Lu, J., et al. (2019). VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems (NeurIPS 2019).
19. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence.
20. Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16(6), 345–379.
21. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
22. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).