Contextual Language Manifold Learning for Operational Trust Quantification in Railway Communication Infrastructures

Syed Akramuddin[1], P. Vamshi Krishna[2*]

[1]PG Student, [2]Assistant Professor, [1,2]Department of Computer Science and Engineering
[1,2]Vaagdevi Engineering College, Warangal, 506005, Telangana, India.
[1]syedsjunnu@gmail.com, [2]vamra1432@gmail.com
*Correspondence: P. Vamshi Krishna (vamra1432@gmail.com)

ABSTRACT

The rapid digitalization of railway communication systems has generated substantial volumes of textual communication records associated with signaling operations, network status, control messages, and security events, making accurate communication security analysis essential for ensuring safe and reliable railway operations. Existing railway communication security assessment primarily relies on manual inspection, expert analysis, and predefined security guidelines to identify secure and non-secure communication records. Although manual analysis supports operational monitoring, it is labor-intensive, time-consuming, difficult to scale with continuously increasing communication data, and often produces inconsistent decisions due to human intervention. These limitations necessitate an intelligent automated framework capable of understanding the semantic context of communication records while providing accurate security classification. To address this need, this study proposes a hybrid semantic learning framework that integrates Sentence Bidirectional Encoder Representations from Transformers (SBERT) with Light Gradient Boosting Machine (LGBM) for railway communication security classification. Initially, communication records undergo preprocessing through text normalization, tokenization, stop-word removal, and lemmatization, after which SBERT generates contextual semantic embeddings that preserve meaningful textual relationships. The extracted feature representations are subsequently classified using the proposed LGBM model, while Tree Alternating Optimization (TAO) Tree, Decision Tree Cost Complexity Pruning (DTCCP), and RuleFit Classifier (RFC) are implemented for comparative performance evaluation. Experimental evaluation using accuracy, precision, recall, F1-score demonstrates the effectiveness of the proposed SBERT-LGBM framework in providing robust semantic representation, efficient classification, and reliable automated security prediction. Furthermore, deployment through a Django-based web application enables scalable, real-time railway communication security analysis, supporting faster decision-making and enhanced protection of critical railway communication infrastructure.

Key words: Railway Communication Security, Natural Language Processing, SBERT, Ensemble Learning, Security Classification, Intelligent Transportation Systems

1. INTRODUCTION

Rail transportation has been one of the most influential modes of public transit since the introduction of the first steam locomotive in the early nineteenth century. Owing to its high passenger capacity, energy efficiency, and lower environmental impact compared with road and air transportation, railways have become an essential component of modern transportation infrastructure. Over the past few decades, rail traffic has experienced substantial growth worldwide, particularly in Europe and Asia, driven by increasing urbanization, economic development, and demand for sustainable mobility solutions. The continuous expansion of rail networks has significantly

Doi:10.48047/ijiee.2026.16.2.53

increased the complexity of railway operations, making safety, reliability, and efficient monitoring critical requirements for railway management systems.

Fig. 1 illustrates the projected growth of the global Railway Cybersecurity Market from 2023 to 2033, categorized into Solutions and Services segments. The market demonstrates a strong and consistent upward trend throughout the forecast period, reflecting the increasing importance of cybersecurity in modern railway transportation systems. In 2023, the market size is approximately USD 7.8 billion, with cybersecurity solutions accounting for the largest share, while services contribute a smaller but steadily growing portion. The market expands gradually to USD 7.9 billion in 2025 and USD 8.5 billion in 2026, driven by the rising adoption of digital rail infrastructure, intelligent signaling systems, IoT-enabled monitoring devices, and cloud-based communication networks. From 2027 onwards, the growth rate accelerates as railway operators invest heavily in advanced threat detection, intrusion prevention, network security, and risk management solutions. By 2033, the railway cybersecurity market is projected to reach approximately USD 17.4 billion, more than doubling its 2023 value. The continuous growth of both solution and service components highlights the increasing need to protect critical railway communication networks, control systems, and operational technologies from cyber threats, ensuring safe, reliable, and resilient rail transportation.

Modern railway systems employ a wide range of sensors and communication technologies to monitor operational conditions and ensure safe train movement. Train bogies, wheel assemblies, braking systems, and communication networks are equipped with multiple sensors that continuously generate large volumes of data. Among these, acoustic, vibration, and communication sensors play a crucial role in identifying abnormal operating conditions and potential equipment failures.

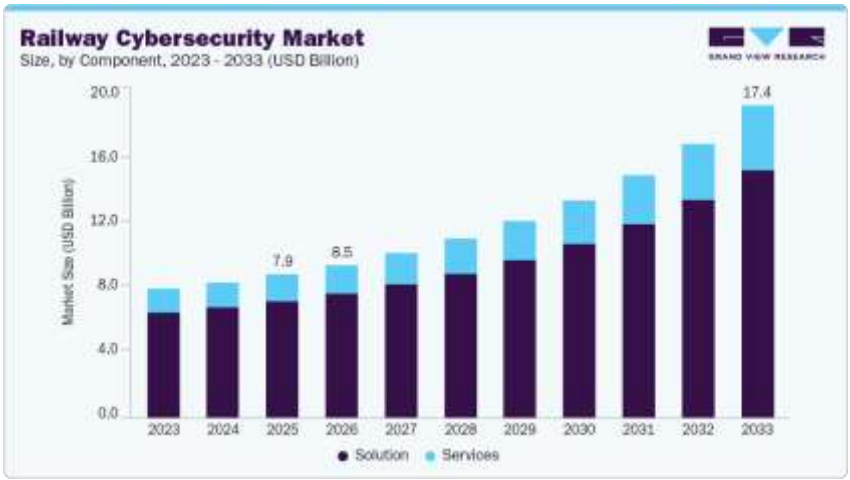

Fig. 1: Global Railway Cybersecurity Market Size by Component (2023–2033).

2. LITERATURE SURVEY

Yang, et al. [6] proposed a transfer learning-based anomaly detection framework using LSTM autoencoders for railway systems. The method first learned normal operational patterns from historical sensor sequences. Transfer learning was then applied to adapt the trained model across different railway operating environments. Reconstruction errors from the autoencoder were utilized to identify abnormal conditions. The approach aimed to reduce retraining requirements when deployment conditions changed. The performance may decrease when the source and target operational domains differ significantly. Zhang, et al. [7] developed TMR-AnoN, a brightness-twin dual-channel network for anomaly detection in railway freight cars. The

Doi:10.48047/ijiee.2026.16.2.53

framework processed complementary image representations through two dedicated channels. Feature extraction modules captured both structural and brightness-related information. The extracted features were fused to improve anomaly discrimination. The network was designed to enhance robustness under varying illumination conditions. The dual-channel architecture increases computational complexity and resource consumption. Zheng, et al. [8] introduced a complete image-based anomaly detection and positioning system for high-speed railway maintenance. The system combined visual inspection techniques with object localization algorithms. Detected anomalies were mapped to their real-world positions using positioning mechanisms. Image processing and detection modules operated in an integrated workflow. The framework supported preventive maintenance activities through automated inspection. Detection accuracy can be affected by environmental factors such as lighting and weather conditions.

Shao, et al. [9] implemented an adaptive transfer learning framework for detecting and predicting anomalous passenger flow in railway stations. The model utilized historical passenger movement data and transferred learned knowledge between stations. Real-time monitoring capabilities enabled rapid anomaly identification. Prediction modules estimated future abnormal passenger flow trends. Adaptive learning mechanisms continuously updated the model with incoming data. Reliable performance depends heavily on the availability of large-scale passenger flow datasets. Liu, et al. [10] proposed a bearing anomaly detection method based on multimodal data fusion and self-adversarial learning. Information from multiple sensing modalities was combined to capture diverse fault characteristics. A self-adversarial learning strategy enhanced feature representation quality. The fused features were used to distinguish normal and faulty bearing conditions. The approach improved robustness against complex operating environments. The requirement for multiple sensor sources increases implementation cost and system complexity. Wang, et al. [11] developed a multi-channel perception-based comfort evaluation system for railway passengers. The framework collected information from various perception channels related to passenger comfort. Multiple indicators were analyzed and integrated into a comprehensive assessment model. Data fusion techniques generated an overall comfort evaluation score. The system supported passenger experience monitoring and improvement. Subjective passenger perceptions may introduce uncertainty into the evaluation process.

Tian and Zou [12] introduced UG-Net, an unsupervised-guided framework for railway foreign object detection. The model learned object representations without relying heavily on labeled datasets. Guidance mechanisms helped distinguish foreign objects from normal railway scenes. Deep feature extraction modules improved object recognition capability. The framework aimed to reduce annotation effort while maintaining detection performance. Unsupervised learning approaches may generate higher false alarm rates than supervised methods. Akçay [13] proposed a hybrid SAINT and CatBoost ensemble model for passenger demand forecasting and anomaly detection. The framework incorporated age-sensitive passenger information into the forecasting process. SAINT captured complex tabular data relationships, while CatBoost enhanced predictive accuracy. Uncertainty estimation mechanisms were used to identify anomalous demand patterns. The ensemble approach improved decision support for urban rail systems. Model interpretability becomes challenging due to the combination of multiple advanced algorithms. Gu, et al. [14] developed an intelligent recognition system for railway vehicle abnormal state detection using YOLOv8. The model employed real-time object detection techniques to identify

Doi:10.48047/ijiee.2026.16.2.53

abnormal vehicle conditions. Feature extraction and localization modules were integrated within a single framework. The system processed visual data collected from railway operations. YOLOv8 enabled fast and accurate recognition of abnormal states. Detection performance may deteriorate when objects are partially occluded or visually degraded.

3. PROPOSED SYSTEM

Thia research was developed to automatically classify railway communication records as Secure or Not Secure by combining semantic feature extraction with machine learning techniques. The framework consists of two phases: a training phase implemented in Jupyter Notebook and a deployment phase implemented using a Django web application. During the training phase, raw railway communication datasets are uploaded, preprocessed, and transformed into semantic feature representations using SBERT, as illustrated in Fig. 2. The generated embeddings are utilized to train multiple classification models, including TAO Tree, DTCCP, RFC, and the proposed LGBM classifier. The trained models are evaluated using standard classification metrics, and the best-performing proposed LGBM model is saved for deployment. During the deployment phase, users upload new communication records through the Django web interface. The uploaded data undergoes the same preprocessing and SBERT-based feature extraction before the saved proposed LGBM model predicts whether each communication record is secure or not secure. The prediction results are then displayed through the web application, enabling railway operators to perform automated, scalable, and real-time communication security monitoring.

.

Data Acquisition and Loading

The proposed framework begins with the collection of railway communication security datasets containing communication messages and related attributes stored in CSV format. During the training phase, the dataset is uploaded into the Jupyter Notebook environment, where it is loaded into memory for further processing. The uploaded dataset serves as the input for preprocessing, feature extraction, model training, and evaluation. During deployment, users upload new communication datasets through the Django web application, allowing the system to perform real-time communication security prediction

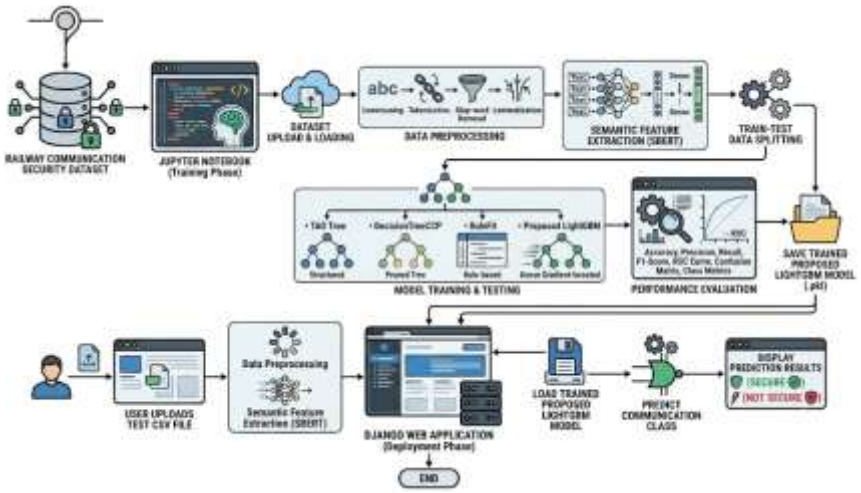

Fig. 2: Proposed system architecture of anomaly detection in rail control communications.

Data Preprocessing

After loading the dataset, the textual communication records undergo preprocessing to improve data quality and eliminate irrelevant information. The preprocessing stage consists of:

Doi:10.48047/ijiee.2026.16.2.53

- Lowercasing all textual content to maintain consistency.
- Tokenizing the communication messages into individual words.
- Removing stop words to eliminate non-informative terms.
- Applying lemmatization to convert words into their root forms.

The processed textual features are combined into a unified representation that serves as input for semantic feature extraction.

Semantic Feature Extraction Using SBERT

Following preprocessing, the cleaned communication records are transformed into contextual semantic embeddings using SBERT. The SBERT model captures the semantic relationships and contextual meaning of railway communication messages by generating dense vector representations. These embeddings preserve both syntactic and semantic information, enabling the classification models to effectively distinguish between secure and insecure communication records. The generated SBERT embeddings are subsequently used for model training during the Jupyter Notebook phase and for prediction during deployment in the Django application.

Model Training and Testing

The extracted SBERT feature vectors are divided into training and testing subsets using an 80:20 train-test split. Multiple machine learning classifiers are trained and evaluated for railway communication security classification. These include:

1. TAO Tree
2. DTCCP
3. RFC
4. Proposed LGBM

Each model learns to classify railway communication records into **Secure** and **Not Secure** categories. After training, the proposed LGBM model demonstrates superior classification performance and is saved as a serialized model (.pkl) for deployment in the web application.

Performance Evaluation

The trained classifiers are evaluated using multiple performance metrics to assess their effectiveness in railway communication security classification. The evaluation includes Accuracy, Precision, Recall, F1-Score, Receiver Operating Characteristic (ROC) Curve, Confusion Matrix, and class-wise performance analysis. Graphical visualizations and performance tables are generated to compare the classification capability of all models and identify the most effective classifier.

Model Deployment and Prediction

The trained proposed LGBM model is integrated into a Django-based web application to provide real-time communication security prediction. Users upload new railway communication datasets through the web interface, after which the uploaded records undergo the same preprocessing and SBERT-based semantic feature extraction used during training. The saved proposed LGBM model then predicts whether each communication record is **Secure** or **Not Secure**. The prediction results are displayed in an interactive table, enabling railway operators to quickly identify potential communication security threats and support timely operational decision-making.

3.1 Proposed LGBM Classifier

The proposed LGBM classifier is an efficient gradient boosting framework designed to perform high-speed and accurate classification by constructing an ensemble of decision trees. Unlike conventional boosting algorithms that grow trees level-wise, LGBM employs a leaf-wise tree growth strategy that expands the leaf with the highest information gain, resulting in improved predictive performance with fewer trees. It also utilizes histogram-based feature binning to reduce memory consumption and computational complexity while maintaining classification accuracy. By combining semantic feature representations generated from SBERT with the learning capability of gradient boosting, the proposed LGBM model effectively captures complex relationships within railway communication records for

Doi:10.48047/ijiee.2026.16.2.53

accurate security classification, as shown in Fig. 3. The model produces reliable predictions while offering faster training and better scalability for large-scale communication datasets.

Input Feature Transformation

The first step involves receiving semantic feature vectors generated by SBERT from the preprocessed railway communication records. Each communication message is represented as a dense embedding vector that preserves contextual and semantic relationships among words. These embeddings serve as the input features for the proposed LGBM classifier, enabling it to learn meaningful patterns associated with secure and non-secure communication records.

Histogram-Based Feature Binning

Instead of evaluating every continuous feature value individually, LGBM groups numerical feature values into discrete histogram bins. This histogram-based representation significantly reduces computational complexity and memory usage while preserving the important characteristics of the original feature distribution. The binned features allow the classifier to efficiently identify optimal split points during tree construction.

Gradient Boosting Tree Construction

Using the histogram-based features, LGBM constructs multiple decision trees sequentially through gradient boosting. Each newly generated tree focuses on correcting the prediction errors produced by the previous trees by minimizing the classification loss. This iterative learning process gradually improves the model's predictive capability and enables it to learn complex non-linear relationships within the semantic feature space.

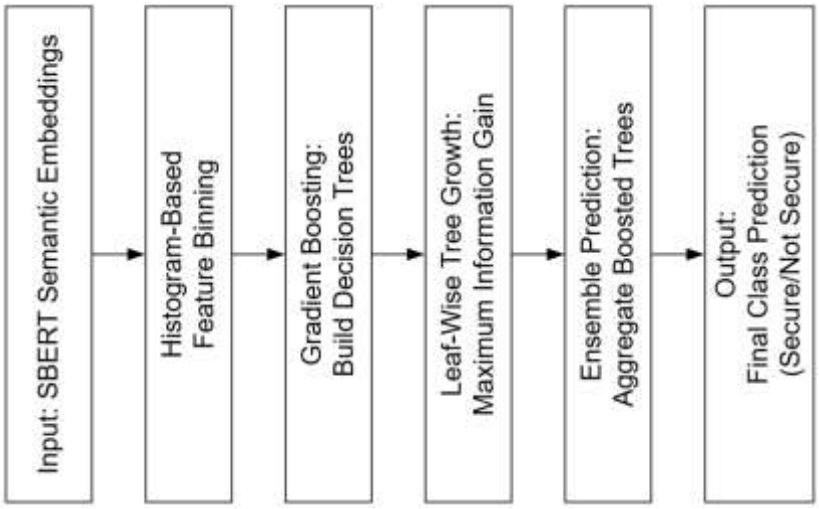

Fig. 3: SBERT-Enhanced LGBM Classifier architecture.

Leaf-Wise Tree Growth

Unlike traditional level-wise tree growth methods, LGBM adopts a leaf-wise growth strategy. During each iteration, the algorithm identifies the leaf node that provides the maximum information gain and expands only that node. This approach produces deeper and more discriminative trees while requiring fewer iterations, resulting in higher classification accuracy and improved learning efficiency.

Final Prediction

After completing the boosting iterations, the outputs of all generated decision trees are aggregated to produce the final classification result. The ensemble prediction determines whether each railway communication record belongs to the "Secure" or "Not Secure" category. The trained proposed LGBM model is subsequently deployed within the Django-based web application to perform real-time communication security prediction on newly uploaded datasets.

Doi:10.48047/ijiee.2026.16.2.53

4. RESULTS AND DISCUSSION

Fig. 4 depicts the confusion matrix of the proposed Paraphrase SBERT-LGBM model, presenting the distribution of correctly and incorrectly classified communication records. The matrix compares the actual class labels with the predicted class labels, enabling detailed evaluation of the model's classification performance. It illustrates the number of True Positives, True Negatives, False Positives, and False Negatives generated during testing, thereby providing insights into the prediction accuracy for each class. The confusion matrix serves as an effective validation tool for measuring the consistency and reliability of the proposed classification framework across different communication categories

.

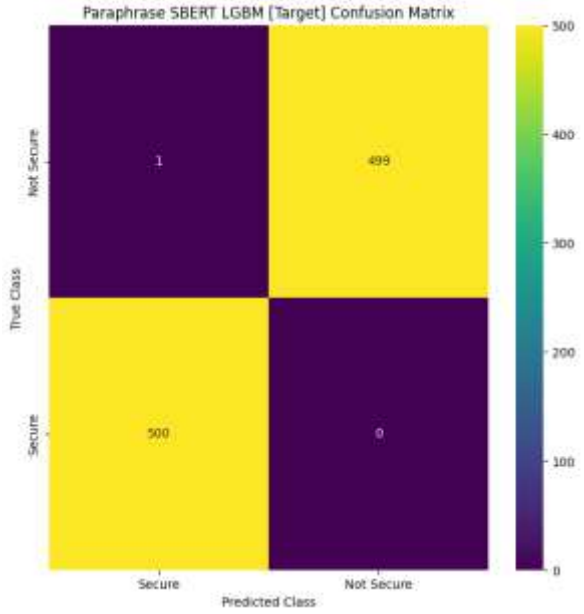

Fig. 4: LGBM Confusion matrix

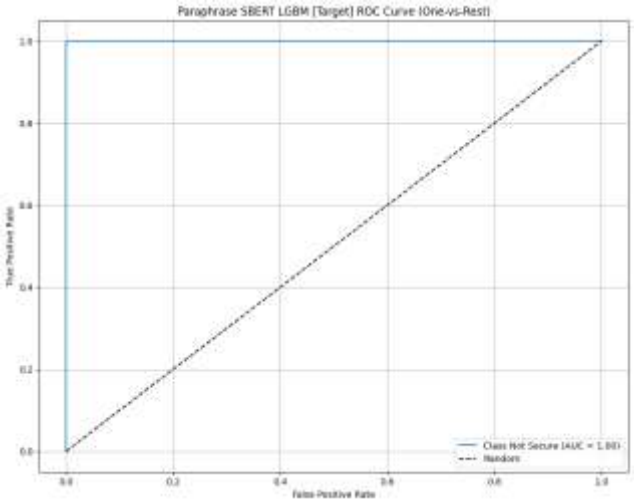

Fig. 5: Paraphrase SBERT LGBM ROC curve.

Fig. 5 illustrates the Receiver Operating Characteristic (ROC) curve of the proposed Paraphrase SBERT-LGBM model for the target classification task. The graph demonstrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR), providing a comprehensive assessment of the classifier's discrimination capability. The ROC

Doi:10.48047/ijiee.2026.16.2.53

curve remains close to the upper-left corner of the graph, indicating that the model effectively distinguishes between the Secure and Not Secure classes with minimal classification error. The Area Under the Curve (AUC) value of 1.00 signifies outstanding classification performance and confirms the robustness of the semantic feature extraction combined with the LGBM classifier.

Fig. 6 illustrates the prediction interface where users can upload a test dataset in CSV format for analysis. Once the dataset is uploaded, the system processes it using the deep learning model to predict potential anomalies in communication data. The results are displayed in a detailed tabular format containing features such as transmission time, error rate, throughput, and data integrity. Each record is analysed, and the final column indicates whether the data is secure or not secure. This feature helps users efficiently evaluate the safety and reliability of railway communication systems based on model predictions.

Fig. 6: Batch Prediction with uploading test data.

Table 1: Performance evaluation obtained using DNDT, DTCCP and proposed RF.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DTCCP [Target]	48.40	24.20	50.00	32.62
TAO Tree [Target]	97.80	97.82	97.78	97.80
RFC [Target]	99.40	99.39	99.41	99.40
LGBM [Target]	99.90	99.90	99.90	99.90

Table 1 presents the comparative performance of the machine learning classifiers used for railway communication security classification. The DTCCP model achieves the lowest performance with an accuracy of 48.40%, indicating limited capability in distinguishing Secure and Not Secure communication records. The TAO Tree classifier significantly improves the classification results, attaining an accuracy of 97.80% with consistently high precision, recall, and F1-score values. RFC further enhances the predictive performance by achieving 99.40% across all evaluation metrics. The proposed Paraphrase SBERT-LGBM model outperforms all comparative classifiers, obtaining 99.90% accuracy, precision, recall, and F1-score. These results demonstrate that the integration of SBERT semantic embeddings with the proposed LGBM classifier provides the most accurate and reliable railway communication security classification among all evaluated models.

Doi:10.48047/ijiee.2026.16.2.53

5. CONCLUSION AND FUTURE SCOPE

This research Paraphrase SBERT-LGBM model represents a significant advancement in railway communication security classification by integrating contextual semantic feature extraction with an efficient gradient boosting framework. By leveraging Paraphrase SBERT's deep contextual language understanding, the model effectively captures complex semantic relationships within railway communication records that conventional machine learning approaches may fail to identify. The proposed model achieves an outstanding 99.90% accuracy, 99.90% precision, 99.90% recall, and 99.90% F1-score, demonstrating exceptional capability in accurately classifying communication records as Secure or Not Secure. The histogram-based learning strategy and leaf-wise tree growth employed by LGBM contribute to improved computational efficiency, faster model training, and superior predictive performance. Furthermore, the integration of SBERT semantic embedding enhances the model's ability to understand contextual communication patterns, resulting in highly reliable classification with minimal misclassification errors. The proposed SBERT-LGBM framework delivers superior accuracy, scalability, and robustness compared with the TAO Tree, DCCP, and RFC classifiers, making it a highly effective solution for intelligent, real-time railway communication security monitoring and decision support systems.

REFERENCES

[1] Jasra, S.K.; Valentino, G.; Muscat, A.; Camilleri, R. A Comparative Study of Unsupervised Deep Learning Methods for Anomaly Detection in Flight Data. Aerospace 2025, 12, 645. https://doi.org/10.3390/aerospace12070645

[2] El-Shafeiy, E.; Alsabaan, M.; Ibrahem, M.I.; Elwahsh, H. Real-Time Anomaly Detection for Water Quality Sensor Monitoring Based on Multivariate Deep Learning Technique. Sensors 2023, 23, 8613. https://doi.org/10.3390/s23208613

[3] Oh, K.; Yoo, M.; Jin, N.; Ko, J.; Seo, J.; Joo, H.; Ko, M. A Review of Deep Learning Applications for Railway Safety. Appl. Sci. 2022, 12, 10572. https://doi.org/10.3390/app122010572

[4] Park, S.; Choi, J.-Y. Hierarchical Anomaly Detection Model for In-Vehicle Networks Using Machine Learning Algorithms. Sensors 2023, 20, 3934. https://doi.org/10.3390/s20143934

[5] Semenov, I.; Swiderski, A.; Borucka, A.; Guzanek, P. Concept of Early Prediction and Identification of Truck Vehicle Failures Supported by In-Vehicle Telematics Platform Based on Abnormality Detection Algorithm. Appl. Sci. 2024, 14, 7191. https://doi.org/10.3390/app14167191

[6] Yang, Chi, Korkut Kaynardag, Junghoon Sohn, and Salvatore Salamone. "Transfer learning in LSTM autoencoders for rail anomaly detection across diverse operational conditions." *Journal of Intelligent Material Systems and Structures* 37, no. 4 (2026): 230-245.

[7] Zhang, Weiyu, Hongmei Shi, Ji Qiu, Jianbo Li, Chao He, and Zujun Yu. "TMR-AnoN: a brightness-twin dual-channel anomaly detection network for railway freight cars." Measurement Science and Technology 37, no. 3 (2026): 036004.

[8] Zheng, Shangdong, Yang Xu, Zhihui Wei, and Zebin Wu. "Image-Based Detection and Real-World Positioning: A Complete System for Anomaly Object Detection in High-Speed Railway Preventive Maintenance." *IEEE Transactions on Instrumentation and Measurement* (2026).

[9] Shao, Yanchun, Enze Liu, Zhiyuan Lin, Shuguang Zhan, and S. C. Wong. "An adaptive transfer learning framework for real-time detection and prediction of anomalous returning passenger flow in railway stations." *Transportmetrica A: Transport Science* (2026): 1-41.

[10] Liu, Han, Yong Qin, and Dilong Tu. "Bearing Anomaly Detection Method Based on Multimodal Fusion and Self-Adversarial Learning." *Sensors* 26, no. 2 (2026): 629.

Doi:10.48047/ijiee.2026.16.2.53

[11] Wang, Zhenyu, Zenan Lu, Zerui Xiang, and Tiecheng Ding. "A multi-channel perception-based comprehensive comfort evaluation system for railway passengers." *Measurement* (2026): 120993.

[12] Tian, Zhuowen, and Jinbai Zou. "UG-Net: An Unsupervised-Guided Framework for Railway Foreign Object Detection." *Applied Sciences* 16, no. 2 (2026): 689.

[13] Akçay, Mehmet Taciddin. "Age-sensitive urban rail passenger demand forecasting and uncertainty-driven anomaly detection using a hybrid SAINT+ CatBoost ensemble." *Scientific Reports* (2026).

[14] Gu, Yibo, Tao Luo, Zhangheng Xu, and Binxin Hu. "Intelligent recognition system for abnormal states of railway vehicles based on YOLOv8." In *International Conference on Frontiers of Traffic and Transportation Engineering (FTTE 2025)*, vol. 14060, pp. 185-192. SPIE, 2026.

Doi:10.48047/ijiee.2026.16.2.53