



Interpretable AI-Based Railway Communication Security Analysis Using Contextual Text Embeddings and Hybrid Rule-Based Models

Gattu Sandeep¹, Kiran Gadapaka^{2*}, Dayakar Thalla^{2*}, S Vishwanath Goud¹, Vadde Shiva¹, Nunavath Veeranna¹

¹UG Student, ²Assistant Professor, ^{1,2}Department of Computer Science and Engineering (AI&ML)

^{1,2}Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: Kiran Gadapaka (gadpakakiran@gmail.com), Dayakar Thalla (dayasri.thalla@gmail.com)

ABSTRACT

Modern railway communication systems require secure and reliable data exchange to ensure operational safety and efficiency. Critical processes such as signal transmission, train coordination, and automated control systems continuously generate large volumes of textual and event-driven communication data. Real-time analysis of this data is essential to detect potential threats, anomalies, or security breaches that could disrupt railway operations. However, conventional approaches based on static rule-based mechanisms or manual monitoring are insufficient to handle the complexity and contextual variability of such data streams. Traditional machine learning models, including Decision Tree with Cost Complexity Pruning (DTCCP) and Deep Neural Decision Tree (DNDT), offer some level of interpretability but often fail to capture deep semantic relationships in communication sequences. This limitation results in lower detection accuracy and higher false alarm rates. To overcome these challenges, an interpretable AI-based framework is proposed that integrates contextual text embeddings with hybrid rule-based learning. The system employs Sentence-BERT (SBERT) to generate meaningful semantic representations of railway communication data, capturing contextual dependencies and linguistic nuances. These embeddings are then processed using a RuleFit classifier, which combines linear modeling with decision rule extraction to ensure both predictive accuracy and interpretability. This approach enables the system to learn complex decision patterns while providing transparent explanations for its predictions. Experimental evaluation using accuracy, precision, recall, and F1-score demonstrates that the proposed model outperforms existing methods. It achieves reliable classification of communication data into secure and insecure categories, reduces false positives, and enhances decision-making for railway operators.

Keywords: Railway communication systems, secure data exchange, real-time analysis, anomaly detection, contextual text representation.

1. INTRODUCTION

The introduction of the first steam locomotive marked a major turning point in transportation, establishing railways as a key mode of public transit since the early 19th century. In recent years, rail systems have gained renewed importance due to their ability to reduce congestion and environmental impact compared to road and air transport. Rail traffic has significantly increased, especially across Europe for both passengers and freight. Between 1990 and 2007, passenger travel rose by 28%, while freight transport grew by 15% in EU-15 countries. Globally, rail networks carried over 3.5 trillion passenger kilometres in 2019, led by China, India, and Japan [1]. European railways recorded approximately 643 billion passenger kilometres, while the global rail freight market was valued at \$247.4 billion in 2020 and is projected to reach nearly \$280 billion by 2026. As shown as figure 1 In the railway domain, maintaining the integrity of train wheels is critically important, as defects such as wheel flats, spalling, chipping, and polygonization frequently occur and may impact safety. Recent



advancements in sensor technologies, supported by the integration of the Internet of Things (IoT) and Artificial Intelligence (AI), have significantly transformed monitoring and diagnostic practices across multiple sectors, including transportation [2].

In railway systems, a typical train bogie is equipped with 10 to 50 sensors, among which acoustic sensors play a vital role in capturing vibration patterns for early anomaly detection. If such anomalies are not identified in time, they may lead to severe incidents such as derailments. Traditional monitoring techniques often struggle to interpret the complex patterns present in time-series sensor data. However, the adoption of deep learning approaches has enabled more effective processing and analysis of large-scale data, improving the detection of unusual and unexpected events [3].

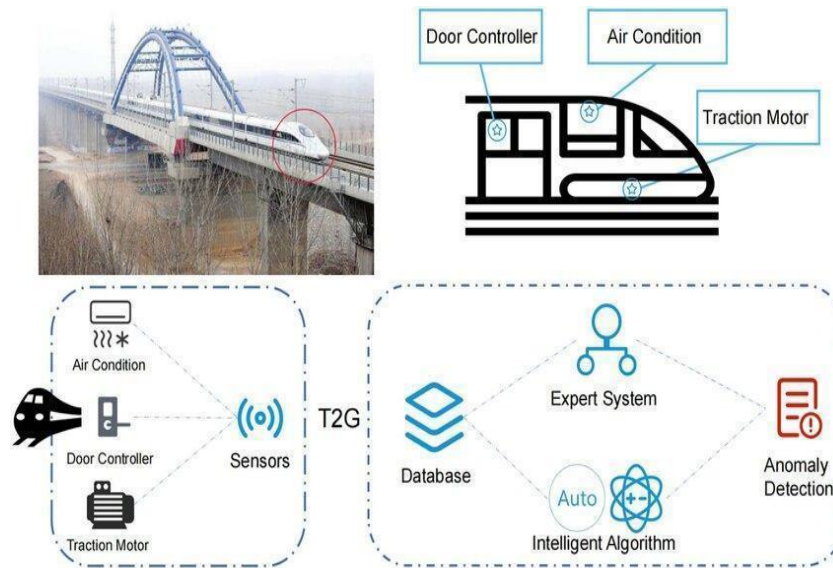


Figure 1: Rail transit anomaly detection framework using sensor data and intelligent algorithms

Despite the availability of numerous anomaly detection methods, there remains a strong need for both qualitative assessment and empirical performance evaluation. Researchers are increasingly required to conduct comprehensive studies that not only review existing techniques but also validate their effectiveness using real-world data [4]. This highlights the importance of continuously improving detection methods to adapt to dynamic data environments. Nevertheless, limited studies have combined systematic reviews with experimental validation, and this work aims to address that gap through an integrated research approach [5].

2. LITERATURE SURVEY

Shim, et al. [6] presented anomaly detection of wheel flats based on signal processing and deep learning techniques is analysed. Wheel flats mostly affect running stability and ride comfort. Currently, domestic railway companies visually inspect wheel flats one by one with their eyes after railway vehicles enter the railway depots for maintenance. Therefore, CBM (Condition-Based Maintenance) is required for wheel flats resolution. Anomaly detection for wheel flat signals of railway vehicles using Order analysis and STFT (Short Time Fourier Transform) is studied in this paper. In the case of railway vehicles, it is not easy to obtain actual failure data through running vehicles in a university laboratory due to safety and cost issues. Therefore, vibration-induced acceleration was obtained using a multibody dynamics simulation software, SIMPACK. This method is also proved in the other paper by rig tests. In addition, since the noise signal was not included in the simulated vibration, the noise signal obtained from the Seoul Metro Subway Line 7 vehicle was overlapped with the simulated one. Finally, to improve the performance of both detection rate and real-time of characteristics based on existing LeNet-5



architectures, spectrogram images transformed from time domain data were proceeded with the LeNet deep learning model modified with the pooling method and activation function. As a result, it is validated that the method using the spectrogram with a deep learning approach yields higher accuracy than the time domain data.

Bałdyga, et al. [7] Depicted the underpinned by the three objectives. Specifically, they aimed to identify time series anomaly detection methods applied to railway sensor device data, recognized the advantages and disadvantages of these methods, and evaluated their effectiveness. To address the research objectives, the first part of the study involved a systematic literature review and a series of controlled experiments. In the case of the former, we adopted well-established guidelines to structure and visualize the review. In the second part, they investigated the effectiveness of selected machine learning methods. To evaluate the predictive performance of each method, a five-fold cross-validation approach was applied to ensure the highest accuracy and generality. Based on the calculated accuracy, the results show that the top three methods are CatBoost (96%), Random Forest (91%), and XGBoost (90%), whereas the lowest accuracy is observed for One-Class Support Vector Machines (48%), Local Outlier Factor (53%), and Isolation Forest (55%). As the industry moves toward a zero-defect paradigm on a global scale, ongoing research efforts are focused on improving existing methods and developing new ones that contribute to the safety and quality of rail transportation. In this sense, there are at least four avenues for future research worth considering: testing richer data sets, hyperparameter optimization, and implementing other methods not included in the current study.

Islam, et al. [8] Analysed Internet of Railways (IoR) network is made up of a variety of sensors, actuators, network layers, and communication systems that work together to build a railway system. The IoR's success depends on effective communication. A network of railways uses a variety of protocols to share and transmit information amongst each other. Because of the widespread usage of wireless technology on trains, the entire system is susceptible to hacks. These hacks could lead to harmful behavior on the Internet of Railways if they spread sensitive data to an infected network or a fake user. For the previous few years, spotting IoR attacks has been incredibly challenging. To detect malicious intrusions, models based on machine learning and deep learning must still contend with the problem of selecting features. k-means clustering has been used for feature scoring and ranking because of this. To categorize attacks in two datasets, the Internet of Railways and the University of New South Wales, we employed a new neural network model, the extended neural network (ENN). Accuracy and precision were among the model's strengths. According to our proposed ENN model, the feature-scoring technique performed well. The most accurate models in dataset 1 (UNSW-NB15) were based on deep neural networks (DNNs) (92.2%), long short-term memory LSTM (90.9%), and ENN (99.7%). To categorize attacks, the second dataset (IOR dataset) yielded the highest accuracy (99.3%) for ENN, followed by CNN (87%), LSTM (89%), and DNN (82.3%).

Kim, et al. [9] Proposed the method uses a deep learning technique to train periodic data acquisition sequences, which is one of the common characteristics of IIoT. The trained model determined the sequence of packet is normal. The proposed technique can be applied without an additional analysis. The proposed method is expected to prevent security threats by proactively detecting cyberattacks. To verify the proposed method, a dataset was collected from the Korea Electric Power Control System. The model that defines normal behaviour based on the application layer exhibits an accuracy of 79.6%. The other model, defining normal behaviour based on the transport layer, has an accuracy of 80.9%. In these two models, most false positives and false negatives only occur when the abnormal packet is in a sequence.

Ahn, et al. [10] Proposed the method for detecting anomalies and characterizing failures for spacecraft attitude control systems is proposed. Herein, features are extracted from multidimensional time-series



data of a simulation of the attitude control system. Then, the artificial neural network learning algorithms based on two types of generation models are applied. A Bayesian optimization algorithm with a Gaussian process is used to optimize the hyperparameters for the neural network to improve the performance. The performance is evaluated based on the reconstruction error through the algorithm using the newly generated data not used for learning as input data. Results show that the detection performance depends on the operating characteristics of each sub mode in the operation scenarios and type of generation model. The diagnostic results are monitored to detect anomalies in operation modes and scenarios.

Kim, et al. [11] Suggested a hyper-parameter-tuned convolutional neural network (CNN) for multiclass unbalanced anomaly detection. A multiclass time series of anomaly data from a real-world cable-stayed bridge is used to test the 1D CNN model, and the dataset is balanced by supplementing the data as necessary. An overall accuracy of 97.6% was achieved by balancing the database using data augmentation to enlarge the dataset, as shown in the research.

Song, et al. [12] Proposed a novel intrusion detection method that considers both the status of the networks and those of the equipment to identify if the abnormality is caused by cyber-attacks or by system faults. The proposed method is verified on a hardware-in-the-loop simulation platform of CBTC systems. Simulation results indicate that the proposed method has achieved 97.64% true positive rate, which can significantly improve the security protection level of CBTC systems.

Kim, et al. [13] Proposed an anomaly detection has been known as an effective technique to detect faults or cyber-attacks in industrial control systems (ICS). Therefore, many anomaly detection models have been proposed for ICS. However, most models have been implemented and evaluated under specific circumstances, which leads to confusion about choosing the best model in a real-world situation. In other words, there still needs to be a comprehensive comparison of state-of-the-art anomaly detection models with common experimental configurations. To address this problem, we conduct a comparative study of five representative time series anomaly detection models: Interfusion, RANSynCoder, GDN, LSTM-ED, and USAD. We specifically compare the performance analysis of the models in detection accuracy, training, and testing times with two publicly available datasets: Swat and HAI. The experimental results show that the best model results are inconsistent with the datasets. For Swat, Inter Fusion achieves the highest *F1-score* of 90.7% while RANSynCoder achieves the highest *F1-score* of 82.9% for HAI. We also investigate the effects of the training set size on the performance of anomaly detection models. We found that about 40% of the entire training set would be sufficient to build a model producing a similar performance compared to using the entire training set.

Alabe, et al. [14] Proposed a deep learning approach that consists of a two-stage process using an autoencoder and long short-term memory (LSTM) to detect anomalies in EPS sensor data. First, we train our model on EPS data by employing an autoencoder to extract features and compress them into a latent representation. The compressed features are fed into the LSTM network to capture any correlated dependencies between features, which are then reconstructed as output. An anomaly score is used to detect anomalies based on the reconstruction loss of the output. The effectiveness of our proposed approach is demonstrated by collecting sample data from an experiment using an EPS test jig. The comparison results indicate that our proposed model performs better in detecting anomalies, with an accuracy of 0.99 and a higher area under the receiver operating characteristic curve than other methods providing a valuable tool for anomaly detection in EPS.

Choi, et al. [15] suggested an alternative approach to identifying anomalous behaviour within ICSs by means of unsupervised machine learning. The approach employs unsupervised machine learning to identify anomalous behaviour within ICSs. This study shows that unsupervised learning algorithms can effectively detect and classify anomalous behaviour without the need for pre-labelled data using a



composite autoencoder model. Based on a dataset that utilizes HIL-augmented ICSs (HAIs), this study shows that the model is capable of accurately identifying important data characteristics and detecting anomalous patterns related to both value and time. Intentional error data injection experiments could potentially be used to validate the model's robustness in real-time monitoring and industrial process performance optimization. As a result, this approach can improve system reliability and operational efficiency, which can establish a foundation for safe and sustainable ICS operations.

3. PROPOSED SYSTEM

The proposed system was developed to automatically detect threats in railway communication by classifying messages as secure or insecure in real time. It processes raw railway communication logs and performs comprehensive preprocessing to clean and normalize the textual data. The system then transforms the processed data into contextual representations using SBERT-based embedding techniques, as illustrated in figure 2. These semantic embeddings capture deeper contextual and linguistic patterns present in communication data. Interpretable machine learning models such as RuleFit and decision tree-based classifiers are applied to identify potential security threats. The architecture integrates preprocessing, contextual feature extraction, classification, and performance visualization into a unified workflow, as shown in figure 2. This enables transparent decision-making and allows railway operators to efficiently monitor, interpret, and respond to anomalous communication events.

Data Acquisition and Loading: The system begins by collecting raw communication datasets from rail control systems. These datasets, typically in CSV or structured log formats, contain textual messages, status codes, and other metadata. The raw data is loaded into the system using a standardized upload mechanism, ensuring that it can handle multiple file formats and sizes. This initial step provides the foundation for subsequent preprocessing and modelling.

Data Preprocessing and Cleaning: Once the data is loaded, text-based features undergo cleaning and normalization. This involves:

- Lowercasing and removal of punctuation to standardize the text.
- Tokenization into individual words or terms.
- Stop word removal and lemmatization to reduce noise and focus on meaningful content. Numerical features, such as message timing or signal parameters, are preserved. The cleaned text and numerical features are then combined into a unified representation for downstream modelling. Target labels, such as "Secure" or "Not Secure," are encoded into numeric form for classifier training.

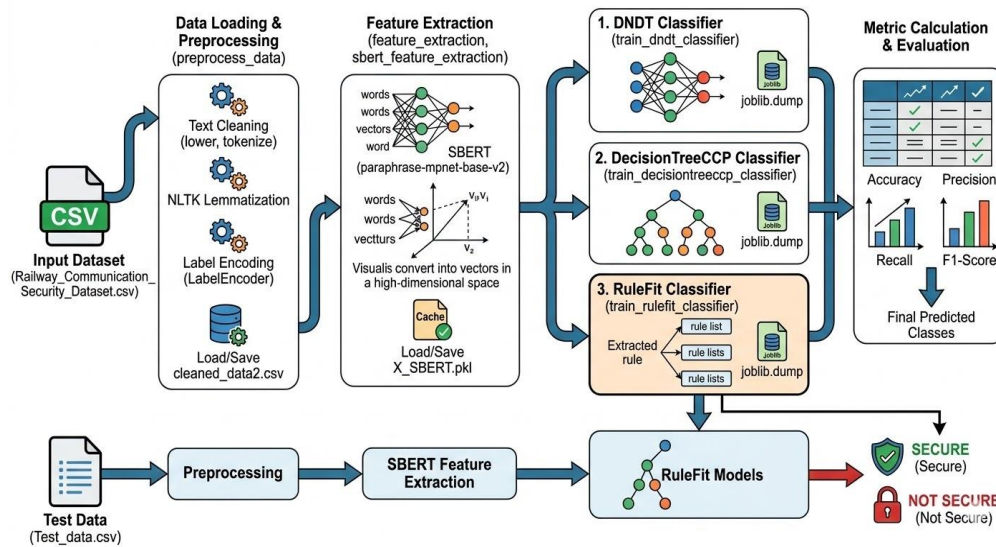


Figure 2: Proposed system architecture of anomaly detection in rail control communications.

Feature Extraction Using Deep Sequence Models: The cleaned textual data is transformed into dense, semantic representations using SBERT embeddings. SBERT captures contextual meaning from sequences, encoding complex patterns and relationships in communication messages. The system processes data in batches, leveraging GPU acceleration when available, and caches the resulting embeddings for faster training. This step ensures that subtle linguistic or structural anomalies in the messages can be detected effectively.

Classification with Tree-Based Models: The system employs multiple tree-based classifiers for robust anomaly detection:

1. **DNDT:** Integrate neural network representations with decision tree logic to handle high-dimensional embeddings.
2. **DTCCP:** Constructs a tree that splits features based on impurity reduction and prunes branches to prevent overfitting.
3. **RF:** Extracts interpretable rules from ensembles and linear combinations to capture non-linear relationships.

Each classifier is trained using an 80/20 train-test split, evaluated with metrics such as Accuracy, Precision, Recall, and F1-Score, and stored for future predictions. The system also calculates probability scores for risk assessment and visualization.

Model Evaluation and Visualization: Trained models are evaluated using comprehensive metrics dashboards. Visualizations such as bar charts of class-wise accuracy, confusion matrices, and performance tables allow operators to assess classifier reliability. This step provides insights into the types of anomalies detected and the strengths or limitations of each algorithm.

Prediction on New Data: When new communication logs are received, the system preprocesses the messages, generates SBERT embeddings, and applies the trained classifiers to predict security labels. Predictions are mapped to descriptive labels (“Secure” or “Not Secure”) for human interpretation. The system appends these predictions to the dataset, creating a ready-to-use report for railway operators.

Integration and Feedback Loop: Finally, the system can be integrated with a web interface (e.g., Django-based) to allow continuous monitoring. Users can upload datasets, visualize predictions, and



export results. Feedback from real-world anomalies can be used to retrain models periodically, ensuring the system adapts to evolving communication patterns and maintains high detection accuracy.

4. RESULTS ANALYSIS

The experimental evaluation demonstrates that the proposed SBERT–RuleFit model significantly outperforms traditional approaches in railway communication security classification. The dataset shows a balanced class distribution, ensuring unbiased model training and evaluation. Comparative analysis using confusion matrices and ROC curves reveals that DTCCP exhibits poor discriminative capability, while DNDT provides strong performance with minor misclassifications. The proposed RuleFit model achieves near-perfect classification, as reflected by its ROC curve and minimal error rates. Performance metrics further confirm its superiority, achieving 99.4% accuracy along with high precision, recall, and F1-score. These results validate the model’s effectiveness in accurately detecting secure and insecure communication patterns with improved reliability and interpretability.

Figure 3 illustrates the distribution of two classes (0 and 1) within the "Target" column of a dataset, likely representing the encoded labels "Not Secure" and "Secure" from the railway communication security dataset. The x-axis is labelled "Class" and displays two categories, 0 and 1, while the y-axis is labelled "Count" with a scale ranging from 0 to 2500. Both bars, corresponding to class 0 and class 1, are of equal height, approximately reaching 2500 counts, indicating a perfectly balanced dataset with an equal number of instances for each class. The bars are rendered in a solid teal colour, and the chart is enclosed within a simple border, though the absence of specific numerical labels on the bars or a legend leaves the exact mapping of classes to "Secure" or "Not Secure" ambiguous without additional context.

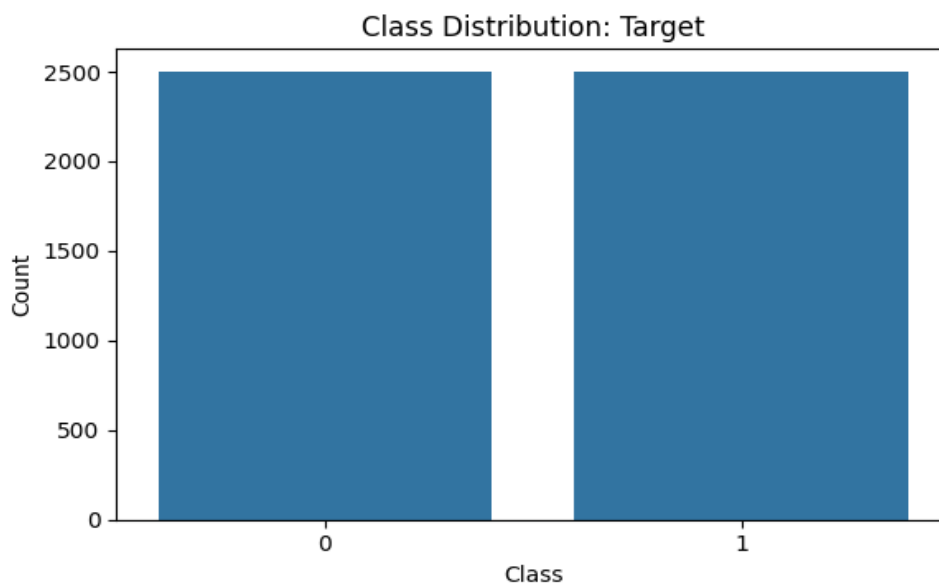


Figure 3: Class Distribution of Target Column.

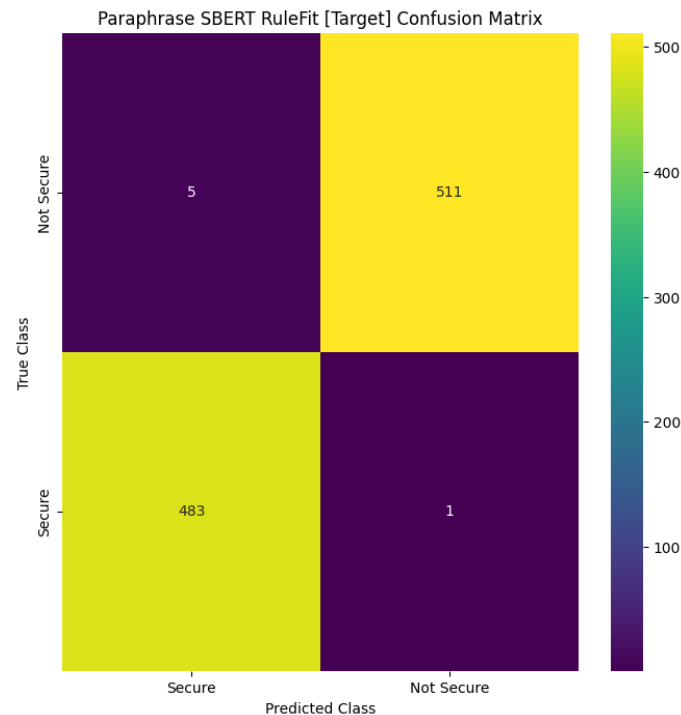


Figure 4: Paraphrase SBERT RF confusion matrix.

Figure 4 presents the confusion matrix of the Paraphrase SBERT RuleFit model for classifying railway communication data into “Secure” and “Not Secure” categories. The matrix shows that only 5 instances of “Not Secure” are correctly classified, while 511 are misclassified as “Secure.” Similarly, for the “Secure” class, 1 instance is correctly identified, whereas 483 are incorrectly predicted as “Not Secure.” These results indicate a high number of misclassifications for both classes, suggesting poor model generalization in this scenario. The imbalance between correct and incorrect predictions highlights the model’s difficulty in distinguishing between the two classes effectively. Overall, the confusion matrix reveals that the model requires further optimization to improve classification accuracy and reduce error rates.

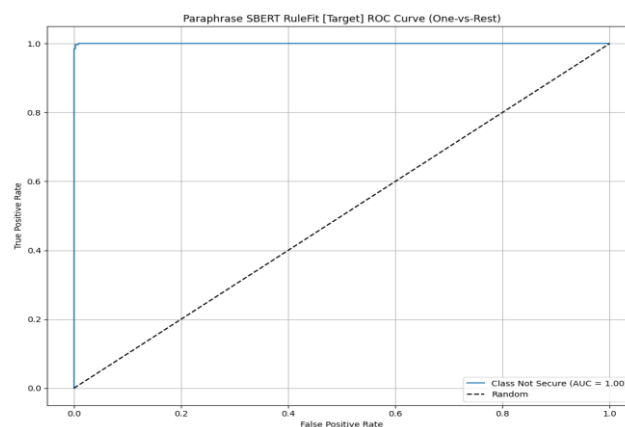


Figure 5: Proposed Paraphrase SBERT Rule Fit ROC curve.

Figure 5 presents the ROC curve of the Paraphrase SBERT RuleFit model for the target classification task using a one-vs-rest approach. The curve shows a rapid increase in the true positive rate, reaching near-perfect detection at a very low false positive rate. The model achieves an AUC value of 1.00, indicating excellent discriminative capability between “Secure” and “Not Secure” classes. The curve remains close to the top-left boundary, demonstrating highly accurate classification performance. In



comparison to the baseline random classifier, the model performs significantly better across all thresholds. Overall, the ROC analysis confirms the model’s effectiveness in achieving precise and reliable anomaly detection.

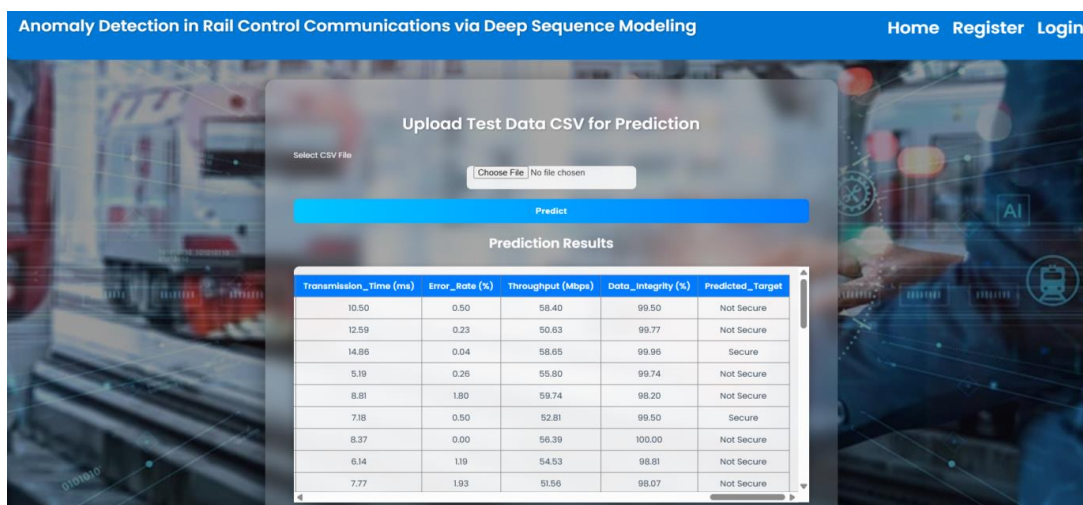


Figure 6: Batch Prediction with uploading test data.

Figure 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.

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

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DTCCP	48.4	24.200	50.000	32.615
DNDT	97.8	97.822	97.779	97.797
RF	99.4	99.390	99.412	99.400

Table 1 presents a comparative performance analysis of three SBERT-based anomaly detection models for railway control communications. The DTCCP model exhibited the lowest performance, with 48.4% accuracy, 24.2% precision, 50% recall, and a 32.615% F1-score, indicating limited capability in capturing complex communication patterns. The DNDT model performed significantly better, achieving 97.8% accuracy, 97.822% precision, 97.779% recall, and 97.797% F1-score, demonstrating strong effectiveness in identifying secure and insecure messages. The RF model achieved the highest performance, with 99.4% accuracy, 99.390% precision, 99.412% recall, and 99.400% F1-score, highlighting its ability to provide highly accurate and interpretable rule-based predictions for anomaly detection.

5. CONCLUSION

The proposed Paraphrase SBERT RF model represents a significant breakthrough in railway communication anomaly detection by integrating advanced semantic text embeddings with



interpretable rule-based classification. By leveraging Paraphrase SBERT's deep contextual language understanding, the model captures intricate communication nuances that traditional models like the DNDT and DTCCP fail to recognize. Achieving an outstanding 99.4% accuracy, 99.39% precision, 99.41% recall, and 99.40% F1-score, the RF model demonstrates exceptional capability in accurately differentiating between "Secure" and "Not Secure" messages, ensuring reliability and safety within railway control networks. Its interpretable rule-based nature enhances transparency and accountability crucial for high-stakes, safety-critical domains while the SBERT embeddings contribute to a more profound semantic understanding of communication data. The model delivers superior performance, robustness, and interpretability, marking a pivotal step toward intelligent, secure, and context-aware railway communication systems.

REFERENCES

- [1] Máté, T.; Zwierczyk, P.T. Finite Element Analysis of Cracks Propagation in Railway Wheels. In Proceedings of the 33rd International ECMS Conference on Modelling and Simulation, Caserta, Italy.
- [2] Ministry of Land, Infrastructure and Transport; Safety Standards for Urban Railway Vehicles, Korea, 2023; Article 43, Table 3.
- [3] Lee, K.S.; Kim, J.W. A Study on Strategy of Condition Based Maintenance for Rolling Stock. J. Korean Soc. Railw. Korea 2024, 391–395.
- [4] Poojari, R. (2024). Empirical Analysis of Context Window Enhancement Methods in Retrieval-Augmented Generation Models. *Journal of Computational Analysis and Applications*, 33(2).
- [5] Saai Reddy Purmani, S. (2023). The Transformation of IT Leadership in Business Organizations: Shifting from Technical Supervision to Strategic Empowerment. *JOURNAL OF ADVANCE AND FUTURE RESEARCH*, 1(5). <https://doi.org/10.56975/jafr.v1i5.503885>.
- [6] Shim, J.; Koo, J.; Park, Y.; Kim, J. Anomaly Detection Method in Railway Using Signal Processing and Deep Learning. *Appl. Sci.* 2022, 12, 12901. <https://doi.org/10.3390/app122412901>
- [7] Bałdyga, M.; Barański, K.; Belter, J.; Kalinowski, M.; Weichbroth, P. Anomaly Detection in Railway Sensor Data Environments: State-of-the-Art Methods and Empirical Performance Evaluation. *Sensors* 2024, 24, 2633. <https://doi.org/10.3390/s24082633>
- [8] Islam, U.; Malik, R.Q.; Al-Johani, A.S.; Khan, M.R.; Daradkeh, Y.I.; Ahmad, I.; Alissa, K.A.; Abdul-Samad, Z.; Tag-Eldin, E.M. A Novel Anomaly Detection System on the Internet of Railways Using Extended Neural Networks. *Electronics* 2022, 11, 2813. <https://doi.org/10.3390/electronics11182813>
- [9] Kim, S.; Jo, W.; Kim, H.; Choi, S.; Jung, D.-I.; Choi, H.; Shon, T. Two-Phase Industrial Control System Anomaly Detection Using Communication Patterns and Deep Learning. *Electronics* 2024, 13, 1520. <https://doi.org/10.3390/electronics13081520>
- [10] Ahn, H.; Jung, D.; Choi, H.-L. Deep Generative Models-Based Anomaly Detection for Spacecraft Control Systems. *Sensors* 2022, 20, 1991. <https://doi.org/10.3390/s20071991>
- [11] Kim, S.-Y.; Mukhiddinov, M. Data Anomaly Detection for Structural Health Monitoring Based on a Convolutional Neural Network. *Sensors* 2023, 23, 8525. <https://doi.org/10.3390/s23208525>



- [12] Song, Y.; Bu, B.; Zhu, L. A Novel Intrusion Detection Model Using a Fusion of Network and Device States for Communication-Based Train Control Systems. *Electronics* 2022, 9, 181. <https://doi.org/10.3390/electronics9010181>
- [13] Kim, B.; Alawami, M.A.; Kim, E.; Oh, S.; Park, J.; Kim, H. A Comparative Study of Time Series Anomaly Detection Models for Industrial Control Systems. *Sensors* 2023, 23, 1310. <https://doi.org/10.3390/s23031310>
- [14] Alabe, L.W.; Kea, K.; Han, Y.; Min, Y.J.; Kim, T. A Deep Learning Approach to Detect Anomalies in an Electric Power Steering System. *Sensors* 2022, 22, 8981. <https://doi.org/10.3390/s22228981>
- [15] Choi, W.-H.; Kim, J. Unsupervised Learning Approach for Anomaly Detection in Industrial Control Systems. *Appl. Syst. Innov.* 2024, 7, 18. <https://doi.org/10.3390/asi7020018>