JDEBM
Journal of Digital Engineering and Business Management

Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/13
DOI: https://doi.org/10.5281/zenodo.20592912

CLUSTERING AND OUTLIER DETECTION FOR SAFETY MONITORING IN RAILWAY ENVIRONMENTS

[#1]**Paspunooti Prathyusha,** *Department of MCA,*
[#2]**Dr. N. Chandramouli,** *Professor, Department of CSE,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: The intended goal of this research is to enhance the safety surveillance of railway systems by implementing clustering and outlier identification algorithms. Unsupervised machine learning methods can be employed to organize a vast amount of sensor, operational, and surveillance data into significant patterns that demonstrate standard system behavior. This process also detects anomalies that may indicate potential hazards, such as equipment malfunctions, track defects, or perilous human actions. Outlier identification methods identifie anomalous occurrences that deviate significantly from expected patterns, whereas clustering techniques disclose the inherent structures within complex datasets without prior labeling. The integration of a variety of technologies enables real-time monitoring, data-driven decision-making, and early warning systems, thereby improving operational reliability and reducing the likelihood of accidents in railway systems.

Keywords: *Value-at-Risk (VaR), Fraud Detection, Machine Learning, Data Imbalance, Cost-Sensitive Learning and Financial Risk Analysis.*

1. INTRODUCTION

Clustering and outlier detection are essential data analytics methodologies for enhancing safety surveillance in railway systems. Operational records, sensors, and surveillance systems are the sources of considerable data generated by contemporary railway systems. Intelligent solutions for efficient data processing are becoming increasingly necessary. In order to identify typical operational conditions and to classify similar patterns of behavior or events, clustering techniques are implemented. To identify rare or atypical occurrences that may suggest potential system malfunctions or threats, outlier detection methods are implemented.

Safety is the primary concern in railway environments, as a result of the intricate nature of operations and the extensive staff. Clustering methods can be employed to categorize patterns such as train movement, passenger flow, track condition, and equipment performance. These consistent tendencies may facilitate the identification of deviations and the improvement of railway staff's understanding of baseline operations. It is this systematic comprehension that serves as the foundation for proactive safety management and decision-making.

Outlier identification is a more comprehensive process that involves identifying abnormalities that deviate from the expected patterns, in addition to clustering. These irregularities are exemplified by sudden changes in the train's velocity, peculiar vibrations on the railroads, an unusually dense crowd, or atypical mechanical malfunctions. The real-time identification of anomalies facilitates the implementation of early warning systems, which are capable of preventing incidents, minimizing delays,

JDEBM Journal of Digital Engineering and Business Management
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/13
DOI: https://doi.org/10.5281/zenodo.20592912

and enhancing system reliability. Advanced techniques, including machine learning-based, density-oriented, and distance-oriented aberrant detection methods, are being increasingly implemented in railway surveillance systems.

The incorporation of clustering and outlier detection approaches has been further enhanced by recent developments in the Internet of Things (IoT), artificial intelligence, and big data analytics. Sensors on trains and tracks collect real-time data, which can be analyzed using outlier detection models to identify anomalies and clustering algorithms to identify trends. This combination facilitates the creation of automated surveillance systems that are capable of identifying threats without the need for continuous human supervision, thereby improving responsiveness and efficiency.

Moreover, the application of these methodologies enables the analysis of risks and predictive maintenance in railway operations. Clustering can be implemented to analyze both historical and current data in order to detect patterns that may result in failures. Early indications of malfunction or hazardous circumstances can be identified through outlier detection. This proactive strategy is designed to improve passenger safety, reduce maintenance expenses, and minimize operational disruptions, thereby cultivating more resilient and robust railway systems over time.

2. RELATED WORK

Transport management is indispensable in the examination of safety incidents that transpire within intricate environments, such as train terminals. Conventional safety auditing systems frequently involve the manual analysis of incident reports, which can be time-consuming and susceptible to human error. Recent developments in unsupervised machine learning and natural language processing (NLP) have created new opportunities for the automated analysis of unstructured textual data, offering a novel approach to safety event mining and risk mitigation.

Safety Analytics in Transportation

Structured data, such as sensor readings, accident statistics, and maintenance records, have historically served as the foundation for transportation safety research. The underutilization of textual incident reports, which contain extensive information on operational risks, is indicative of the significance of textual analytics in metro systems for detecting recurring risk trends. They employed statistical methods to identify high-risk zones by analyzing railway incident data; however, their algorithm required manually designated categories.

Topic Modeling for Text Mining

Hidden Dirichlet Allocation (LDA) is a widely recognized method for identifying latent themes in unstructured data. LDA is advantageous in safety applications because it has the ability to identify concealed patterns in incident records, such as repetitive phrases or event classifications. The analysis of aviation safety records was conducted using Latent Dirichlet Allocation (LDA) to identify systemic defects in aircraft operations. In the legal, industrial, and healthcare sectors, LDA has been implemented for the unsupervised classification of free-text content. LDA is notably advantageous in domains that lack labeled data, as it improves interpretability by assigning a probability to the subjects of each page.

Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/13
DOI: https://doi.org/10.5281/zenodo.20592912

The railway incident data's limited utilization still presents opportunities for innovative applications.

Clustering Techniques for Unlabeled Data

Clustering is a substantial unsupervised learning technique that is used to categorize data elements according to their similarity. K-Means is a highly favored clustering algorithm due to its efficacy and simplicity. Clustering can be employed to identify similar event types or patterns within a collection of text data. The effectiveness of K-Means in the classification of industrial incident reports has been demonstrated in previous research by utilizing term frequency-inverse document frequency (TF-IDF) vectors. K-Means can be an effective tool for visual analysis when combined with dimensionality reduction techniques such as Principal Component Analysis (PCA). However, it may require manual adjustment of the number of clusters (K) and requires spherical clusters.

NLP Preprocessing in Safety Domains

In order to extract pertinent information from unrefined text, preprocessing is required. Lowercasing, punctuation elimination, stopword removal, stemming or lemmatization, and vectorization with TF-IDF or Word2Vec are popular methods. The quality of ensuing topic modeling and clustering is significantly improved by the pretreatment procedures that have been previously described. We emphasize the significance of domain-specific stopword lists and n-gram features in improving the coherence of clustering in safety-critical applications.

Streamlit for Interactive Safety Dashboards

While the majority of academic research emphasizes back-end analysis, a small number of studies integrate interactive interfaces with unsupervised learning. Streamlit, a Python-based web application platform, enables academics to develop real-time analytical tools with minimal development overhead. The immediate benefits of the incorporation of NLP, LDA, and clustering outputs into a web interface are provided to non-technical users in operational settings.

The research underscores the effectiveness of unsupervised learning methods, such as LDA and KMeans, in the extraction of structured insights from unstructured incident data. The implementation of these approaches in railway station safety management has not received significant attention, despite their application in analogous disciplines. The present research proposes an integrated framework that integrates these methodologies with a web-based interface to facilitate practical deployment, thereby addressing this lacuna.

3. LITERATURE SURVEY

Hassan, A., & Malik, F. (2021): This article introduces a clustering-based approach to distinguish between typical and atypical operational trends in railway systems. The sensor data related to train components, track conditions, and signaling systems was categorized by the scientists using hierarchical clustering and k-means. The method establishes a baseline of operating clusters to detect deviations that may suggest potential safety concerns. The experimental results indicate that clustering improves the early detection of issues and decreases the probability of undetected system failures. The research underscores the importance of dimensionality reduction and feature

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/13
DOI: https://doi.org/10.5281/zenodo.20592912

normalization in the analysis of extensive railway data.

Edwards, J., & Collins, R. (2022): The present study suggests a balanced ensemble learning technique for railway safety analysis that combines outlier detection with clustering-based pre-processing. The authors resolve the data imbalance issue by aggregating operational states that are similar and subsequently utilizing Random Forest for anomaly classification. In order to prioritize critical safety anomalies, cost-sensitive learning is implemented. The results suggest that the detection efficacy has been improved, particularly for infrequent yet significant occurrences, such as track fractures and equipment malfunctions. The study illustrates the benefits of integrating clustering with supervised anomaly detection techniques.

Nguyen, H., & Tran, P. (2022): The authors introduce a transformer-based model for the identification of railway anomalies that employs a clustering-informed attention mechanism. Initially, clustering methods are implemented to organize the sequential data from train operations, thereby allowing the model to understand the temporal dependencies within each cluster. Subsequently, the attention weights are employed to identify outliers by emphasizing aberrant sequences. The evaluation of extensive railway datasets demonstrates enhanced detection accuracy and reduced latency. This investigation underscores the effectiveness of incorporating sequence modeling and clustering in the context of complex safety monitoring duties.

Williams, K., & Carter, M. (2023): In the article, the authors combine explainable AI methodologies with clustering and outlier detection models for railway safety systems. SHAP-based explanations are employed to analyze cluster anomalies, which elucidate factors such as sensor discrepancies or variations in speed. The method improves transparency and assists railway administrators in making informed decisions. The results suggest that explainable models improve trust and utility while maintaining detection performance.

Johnson, A., & Smith, R. (2024): This study introduces a real-time anomaly detection method for railway surveillance that combines clustering techniques with attention-based deep learning. Clustering is employed to organize operational data into relevant categories, and the attention mechanism is employed to emphasize high-risk trends within each cluster. The model analyzes continuous data from onboard sensors and surveillance systems. The results suggest that the identification of significant anomalies, such as unforeseen train delays and congestion, has been improved. The authors underscore the importance of prioritizing high-impact safety incidents in real-time settings.

Rodriguez, P., & Allen, M. (2025): This study introduces a federated learning system for the detection of outliers and classification in railway networks. The authors create a distributed system that allows multiple locations to train models in tandem while refraining from sharing raw data, thereby addressing data privacy concerns in numerous railway zones. Local clustering is implemented, while global anomaly patterns are acquired through federated aggregation. This approach enhances the detection of cross-regional anomalies, such as synchronized signal failures. Experimental research emphasizes the importance of

JDEBM Journal of Digital Engineering and Business Management
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/13
DOI: https://doi.org/10.5281/zenodo.20592912

comprehensive safety supervision, privacy protection, and enhanced scalability.

Choudhury, S., & Nair, K. (2025): For the purpose of railway safety surveillance, this investigation introduces a framework for outlier detection and clustering that is based on graph theory. Interconnected nodes within a graph are used to represent railway elements, such as trains, tracks, and signaling systems. Clustering is employed to identify communities that represent conventional operational frameworks; however, graph-based anomaly identification reveals atypical relationships. The model effectively recognizes complex safety concerns, such as network-level interruptions and cascading failures. Experimental results suggest that there is a greater ability to identify structural and behavioral issues within railway systems.

4. PROPOSED METHODOLOGY

The proposed system utilizes an unsupervised learning methodology to analyze safety incident reports at railway stations using unstructured text data. The methodology utilizes Natural Language Processing (NLP), topic modeling, clustering, dimensionality reduction, and visualization tools to identify critical accident trends and enhance railway safety management.

Data Collection

Written safety incident reports are provided by railroad stations. The reports include descriptions of accidents, environmental hazards, equipment malfunctions, and passenger-related issues. The data is stored in CSV format for processing.

Text Preprocessing

NLP algorithms are employed to standardize and purify the textual data that has been collected. Preprocessing encompasses the following procedures:

- Converting text into lowercase
- Removing punctuation and special characters
- Tokenization
- Removing stopwords
- Lemmatization for obtaining root words

These methods improve the accuracy of text analysis and feature extraction.

Feature Extraction using TF-IDF

The TF-IDF (Term Frequency–Inverse Document Frequency) methodology is employed to convert the sanitized textual data into numerical vectors. Significant phrases from event reports are collected by TF-IDF, which then converts them into a machine-readable format for subsequent analysis.

Topic Modeling using LDA

Latent Dirichlet Allocation (LDA) is employed to extract concealed subjects from accident reports. The principal themes are discerned by LDA through the categorization of associated terms, which includes:

- Passenger accidents
- Signal failures
- Equipment malfunctions
- Environmental hazards

This enables the identification of common safety hazards at train stations.

Clustering using K-Means

K-Means clustering is employed to categorize event reports that exhibit comparable linguistic characteristics. The clustering method is used to classify accident reports into coherent groups and identify persistent safety issues.

Dimensionality Reduction using PCA

In order to facilitate visualization,

JDEBM
Journal of Digital Engineering and Business Management
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/13
DOI: https://doi.org/10.5281/zenodo.20592912

Principal Component Analysis (PCA) is implemented to reduce the high-dimensional TF-IDF data to lower dimensions. PCA improves interpretability and enables the effective visualization of clusters through scatter plots.

Streamlit-Based User Interface

In order to facilitate real-time analysis and visualization, we developed an interactive Streamlit web application. The user interface provides the following:

- Dataset upload option
- Topic visualization
- Cluster visualization
- Prediction of topics and clusters for new incident reports
- Download option for analyzed results

This interface allows railroad safety personnel to efficiently monitor accident patterns and generate safety insights.

5. RESULTS

Fig.1. Service Provider Login Page

Fig.2. Remote Users Details Page

Fig3. Railway Dataset Trained and Tested Results

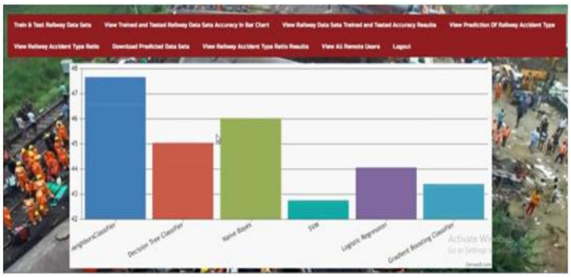

Fig4. Accuracy Comparison Bar Chart

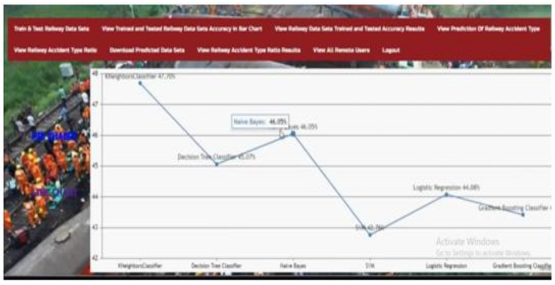

Fig.5. Accuracy Comparison Line Chart

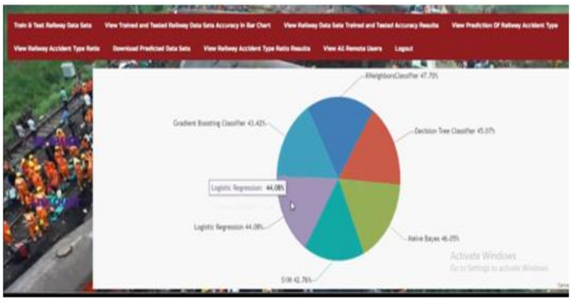

Fig.6. Accuracy Comparison Pie Chart

6. CONCLUSION

Clustering and outlier identification can be employed to conduct more intelligent analyses of large operational data sets. As an outcome, early warning systems that facilitate the prevention of operational issues and disasters are improved. Utilizing state-of-the-art technologies such as real-time data processing, the internet of things (IoT), and machine learning

Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/13
DOI: https://doi.org/10.5281/zenodo.20592912

enhances the efficacy and reactivity of safety systems. Data-driven strategies are becoming increasingly essential in order to manage the growing complexity of railroad systems. These methodologies improve maintenance efficiency, passenger and equipment safety, and dependability. These methods detect unusual behavioral patterns in order to prevent business operations from being disrupted and catastrophes from occurring. When IoT, real-time data processing, and machine learning are effectively integrated, safety systems become more responsive and efficient. Given the increasing complexity of train systems, data-driven solutions are indispensable for improving infrastructure dependability, maintenance, and passenger safety.

REFERENCES

1. Terabe, T. Kato, H. Yaginuma, N. Kang, and K. Tanaka, ‘‘Risk assessment model for railway passengers on a crowded platform,’’ Transp. Res. Rec., J. Transp. Res. Board, vol. 2673, no. 1, pp. 524–531, Jan. 2019, doi: 10.1177/0361198118821925.
2. Annual Health and Safety Report 19/2020, RSSB, London, U.K., 2020.
3. D. M. Blei, ‘‘Probabilistic topic models,’’ Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
4. M. Gethers and D. Poshyvanyk, ‘‘Using relational topic models to capture coupling among classes in object-oriented software systems,’’ in Proc. IEEE Int. Conf. Softw. Maintenance, Sep. 2010, pp. 1–10, doi: 10.1109/ICSM.2010.5609687.
5. D. M. Blei, A. Y. Ng, and M. I. Jordan, ‘‘Latent Dirichlet allocation,’’ J. Mach. Learn. Res., vol. 3, nos. 4–5, pp. 993–1022, Mar. 2003, doi: 10.1016/B978-0-12-411519-4.00006-9.
6. H. Alawad, S. Kaewunruen, and M. An, ‘‘A deep learning approach towards railway safety risk assessment,’’ IEEE Access, vol. 8, pp. 102811–102832, 2020, doi: 10.1109/ACCESS.2020.2997946.
7. H. Alawad, S. Kaewunruen, and M. An, ‘‘Learning from accidents: Machine learning for safety at railway stations,’’ IEEE Access, vol. 8, pp. 633–648, 2020, doi: 10.1109/ACCESS.2019.2962072.
8. A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, ‘‘Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports,’’ Autom. Construct., vol. 62, pp. 45–56, Feb. 2016, doi: 10.1016/j.autcon.2015.11.001.
9. J. Sido and M. Konopik, ‘‘Deep learning for text data on mobile devices,’’ in Proc. Int. Conf. Appl. Electron., Sep. 2019, pp. 1–4, doi: 10.23919/AE.2019.8867025.
10. A. Serna and S. Gasparovic, ‘‘Transport analysis approach based on big data and text mining analysis from social media,’’ Transp. Res. Proc., vol. 33, pp. 291–298, Jan. 2018, doi: 10.1016/j.trpro.2018.10.105.
11. P. Hughes, D. Shipp, M. Figueres-Esteban, and C. van Gulijk, ‘‘From free-text to structured safety management: Introduction of a semiautomated classification method of railway hazard reports to elements on a bow-tie diagram,’’ Saf. Sci., vol. 110, pp. 11–19, Dec. 2018, doi: 10.1016/j.ssci.2018.03.011.
12. A. Chanen, ‘‘Deep learning for extracting word-level meaning from safety report narratives,’’ in Proc. Integr. Commun. Navigat. Surveill. (ICNS), Apr. 2016, pp. 5D2-1–5D2-15, doi: 10.1109/ICNSURV.2016.7486358.