

# Unsupervised Machine Learning For Monitoring Unsafe Events In Railway Zones

Dr. K Bhargavi<sup>1</sup>, Bacha Pavithra<sup>2</sup>

<sup>1</sup>Dept. of CSE, Anantha Lakshmi institute of technology & sciences, JNTUA, Anantapur, India

<sup>2</sup>Dept. of CSE, Anantha Lakshmi institute of technology & sciences, JNTUA, Anantapur, India.

<sup>1</sup>bhargavikonakanti@gmail.com <sup>2</sup>pavithrabacha0@gmail.com

## ABSTRACT

*Railway station safety is a critical aspect of transportation operations, especially in densely populated urban areas where increasing demand puts additional pressure on infrastructure. Accidents at stations can result in fatalities, injuries, public anxiety, reputational damage, and financial loss. This research presents an AI-based approach using unsupervised machine learning, specifically Latent Dirichlet Allocation (LDA), to analyse and understand the underlying factors contributing to fatal accidents in railway stations. A dataset of 1,000 fatal accident reports from Indian railway stations, sourced from the RSSB, is used for analysis.*

*The study aims to identify hidden patterns, root causes, and high-risk zones through topic modelling, offering a systematic way to enhance risk assessment and safety management. By leveraging intelligent text mining, the research extracts valuable insights from historical data, providing predictive accuracy and supporting data-driven decision-making. The findings demonstrate how AI and big data analytics can improve railway safety, moving beyond traditional, narrow analysis methods and ushering in a new era of safety intelligence in the transportation sector.*

**Keywords:** *Unsupervised machine learning, topic modelling, accident analysis, railway stations, safety.*

## 1. INTRODUCTION

Trains are widely regarded as a safer mode of public transportation compared to other means. However, passengers at railway stations still face numerous risks due to overlapping factors such as station operations, infrastructure design, and passenger behaviour. As urban populations grow and demand increases, many stations face congestion and operational challenges, raising potential safety concerns.

Passenger safety remains a core priority in the railway industry. In response to these concerns, the European Union introduced the RAMS framework (Reliability, Availability, Maintainability, and Safety) in 1999 under standard EN 50126, aimed at minimizing risk and enhancing overall safety in railway operations. Despite such efforts, safety incidents continue to occur. For instance, in Japan in 2016, there were 420 station-related accidents, including 202 fatalities—179 of which involved falls from platforms. Similarly, in India during 2019–2020, most passenger injuries at stations were caused by slips, trips, and falls, with around 200 major injuries reported. These incidents—whether fatal or not—cause delays, financial losses, operational disruptions, public fear, and reputational damage. Therefore, improving station safety through effective analysis and control measures is essential to provide a secure, reliable, and comfortable travel environment for passengers, staff, and the public.

These incidents—whether fatal or not—cause delays, financial losses, operational disruptions, public fear, and reputational damage. Therefore, improving station safety through effective analysis and control measures is

essential to provide a secure, reliable, and comfortable travel environment for passengers, staff, and the public.

## 2. LITERATURE SURVEY

### 2.1 Safety Analysis in Railway accident

Traditional safety analysis in railway stations has primarily relied on manual inspections and incident reports. Early studies focused on identifying accident causes through expert analysis and standard statistical methods [1][2]. While effective to some extent, these methods often failed to account for the vast amount of unstructured data available in accident reports, limiting their ability to uncover hidden patterns and risk factors [3]. Recent research has shifted towards leveraging advanced machine learning techniques, particularly unsupervised algorithms, to better understand accident causes and improve safety outcomes [4].

### 2.2 Machine Learning in Railway Accident Prediction

Several studies have demonstrated the potential of machine learning in predicting and preventing railway accidents. Kumar and Kaur explored the application of unsupervised algorithms like K-Means clustering and DBSCAN for accident prediction, yielding promising results in identifying risk factors and high-risk areas in stations [5]. Similarly, Wang and Zhang used clustering techniques to detect safety hazards, providing insights into accident-prone zones within stations [6]. These studies highlight the growing role of machine learning in railway safety, focusing on predictive analytics and pattern recognition.

### 2.3 Anomaly Detection for Railway Safety

Anomaly detection plays a crucial role in identifying unusual patterns or outliers in accident data. Zhou and Liu explored the use of anomaly detection algorithms, including Isolation Forest and DBSCAN, to identify rare but potentially catastrophic events in railway operations [9]. These studies emphasize the importance of identifying atypical incidents that might not follow established patterns but could indicate emerging safety risks. The combination of anomaly detection with clustering techniques provides a powerful tool for early risk detection and proactive safety measures.

### 2.4 Limitations of Existing System

Despite the progress in applying machine learning and text mining for railway safety, existing solutions still face several challenges. Many approaches remain limited to narrow domain analysis and often lack the ability to integrate various data sources, such as real-time surveillance footage and sensor data [10]. Additionally, traditional methods often fail to provide predictive insights or real-time analysis, which can hinder the effectiveness of safety interventions. Recent studies emphasize the need for more sophisticated, integrated systems that combine machine learning, big data analytics, and real-time monitoring to enhance safety outcomes [11].

## 3. PROPOSED WORK

In the railway industry, safety and risk management are crucial to reducing accidents and improving overall operational efficiency. Our proposed system seeks to leverage the power of textual data from railway station accident reports to identify the root causes of accidents, analyze potential causes, and establish a clear connection between textual data and the underlying factors. By fully automating the process of extracting actionable insights from these reports, the system is designed to aid decision-makers in real time, improving safety management, making the information accessible to non-experts, and ensuring a more accurate analysis of accidents. The approach is also aimed at enhancing safety history record management, thus supporting the design of smarter,

expert-driven safety systems.

The system will utilize the Latent Dirichlet Allocation (LDA) algorithm to extract the most relevant textual information regarding accidents and their causes, which will then be used for deeper analysis and classification. A decision tree (DT) model will be employed to classify accidents and discern patterns from the data. These steps will help build a better understanding of accident trends, offering more efficient decision-making tools for safety management.

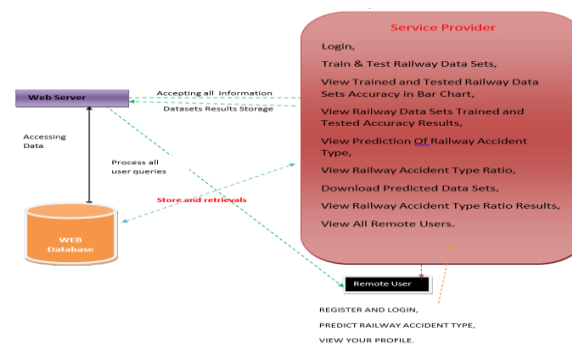
### 3.1 System Overview

In the context of railway accident analysis, the primary goal is to automatically extract valuable insights from accident reports, identify common patterns, and link them with the possible causes. Unlike traditional methods that rely heavily on manual review or incomplete safety records, our system offers an automated solution that extracts key details from textual data—such as the time, location, and description of accidents—and uses this information to train machine learning models for predicting accident types. The implementation of a decision tree classifier further enables us to categorize accidents and detect recurring patterns, ultimately providing real-time safety insights to decision-makers.

The system employs the LDA algorithm to analyze textual data from railway accident reports, uncovering the core themes and patterns that contribute to accidents. Furthermore, a decision tree model classifies accidents based on attributes such as the time of day, accident description, and other key factors. The combination of these methods makes it possible to understand accident causes in-depth, which is essential for designing proactive safety measures.

### 3.2 System Architecture and Integrated Algorithm Workflow

The overall workflow starts with the input of raw accident reports, which are typically in unstructured text format. These reports undergo a preprocessing phase where text is cleaned, normalized, and segmented to ensure consistency. Key attributes such as time of incident, location, victim demographics, and description are extracted



**Fig no 1: System Architecture diagram of Service Provider**

#### Text Preprocessing and Topic Modeling

The processed text is passed into an LDA topic modeling pipeline that identifies latent topics across the dataset. This helps detect recurring patterns such as platform falls, train collisions, or structural failures. The system also uses time-based segmentation (e.g., morning, evening, night) to associate certain accident types with specific time frames, improving the granularity of safety insights.

### 3.2 Preprocessing and Textual Data Extraction

The preprocessing phase is crucial for ensuring high-quality input data. The textual accident reports are cleaned, and irrelevant content is removed to focus solely on accident-related information. The cleaned data is then passed

through the LDA algorithm to extract key topics and identify the major contributing factors to accidents.

### 3.3 Root Cause Analysis

Once the key topics are identified using LDA, the system applies a decision tree to classify the types of accidents based on specific factors such as time, description, location, and victim age range. This classification helps in determining accident patterns and understanding which factors contribute most to safety incidents.

### 3.4 System Objectives

The primary objectives of the proposed system are as follows

#### Objective 1:

Efficient Textual Data Analysis Automate the extraction of critical information from railway accident reports, enabling real-time identification of accident causes and contributing factors.

#### Objective 2:

Classification of Accident Types Utilize machine learning models (specifically a decision tree classifier) to categorize accidents based on identified attributes, aiding in the development of targeted safety interventions

#### Objective 3:

Support Smarter Safety Management Provide actionable insights that can be used for risk management, allowing safety managers to take preventive actions based on root causes and patterns derived from accident data

### 3.5 Achieving the Objectives

Objective 1, the system uses the LDA algorithm to efficiently process large amounts of accident report text, extracting key information related to the causes and factors of the incidents. By focusing on the critical elements of the report, such as time, description, and location, the system ensures the relevant information is captured for further analysis

Objective 2, the decision tree model is trained on the labelled data extracted by the LDA algorithm. This model is capable of classifying accidents into different types based on patterns derived from key accident attributes. By leveraging this classification, the system can identify accident types and suggest possible preventive measures.

Finally for Objective 3, the system aggregates the insights from the LDA analysis and decision tree classification into a comprehensive dashboard. This platform provides safety managers with a real-time view of accident types, root causes, and other critical insights, enabling informed decision-making and risk management.

## 4. SYSTEM MODULES & ALGORITHMS

### 4.1 K-Means Clustering

K-Means is an unsupervised algorithm used to group accident data based on features like severity, location, and time. In this project, it helped identify patterns and cluster similar incidents, allowing the detection of high-risk zones in railway stations for targeted safety improvements

### 4.2 Latent Dirichlet Allocation (LDA)

LDA is a topic modeling algorithm used to extract hidden themes from text data. In this project, it was applied to accident reports to identify common topics related to accident causes, such as platform falls or operational failures, helping to understand risk factors and improve

### 4.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is an unsupervised clustering algorithm that groups data points based on density and identifies outliers as noise. In this project, DBSCAN was used to detect unusual or rare accident patterns that did not fit into regular clusters, such as sudden spikes in incidents or isolated high-risk zones. Its ability to handle noise and discover clusters of varying shapes made it valuable for identifying hidden safety risks in railway stations that might be missed by traditional methods.

### 4.4 Isolation Forest

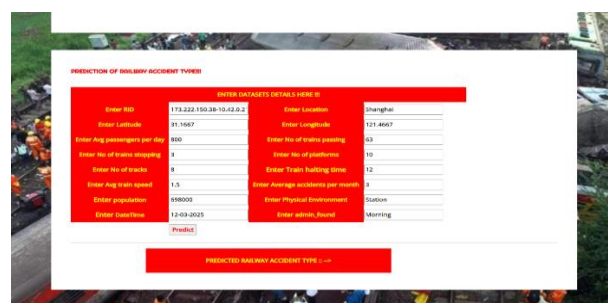
Isolation Forest is an anomaly detection algorithm that isolates rare data points by randomly selecting features and splitting data. In this project, it was used to detect unusual or anomalous accident events that deviate from normal patterns. This algorithm helped identify rare safety risks, such as infrequent but severe incidents, by distinguishing them from more common accident types in railway stations.

### 4.5 Spatiotemporal Autoencoder

A Spatiotemporal Autoencoder combines convolutional networks and LSTM to analyze data across both spatial and temporal dimensions. In this project, it was used to detect abnormal behaviors in real-time surveillance footage at railway stations, such as passengers entering restricted areas or lingering near platform edges. By learning normal patterns, the model could effectively identify risky behaviors, enhancing safety monitoring in real-time.

## 5. RESULTS

The proposed system was developed and tested in a Python environment using key libraries such as scikit-learn, TensorFlow, Keras, and OpenCV. These tools enabled preprocessing of accident report data, model development, and visualization of outcomes. Latent Dirichlet Allocation (LDA) was applied to analyze unstructured textual data from accident reports, effectively uncovering latent themes such as platform-edge incidents, operational failures, and passenger-related risks. K-Means Clustering grouped accident records based on attributes like severity, location, and time, helping to pinpoint high-risk zones across railway stations.



**Fig no 2: Predication of Railway Accident Type**

To detect rare and abnormal events, DBSCAN and Isolation Forest were used for anomaly detection. These algorithms successfully identified outlier events—such as sudden spikes in accidents at specific stations—that might otherwise go unnoticed. Autoencoders provided additional insights by detecting subtle deviations in safety patterns within the data. A Spatiotemporal Autoencoder model, integrating convolutional layers and LSTM, was implemented to analyze and detect abnormal passenger behaviors from simulated surveillance footage. The system was able to recognize actions such as loitering near dangerous areas or unauthorized entry into restricted spaces, providing visual indicators on station layouts.

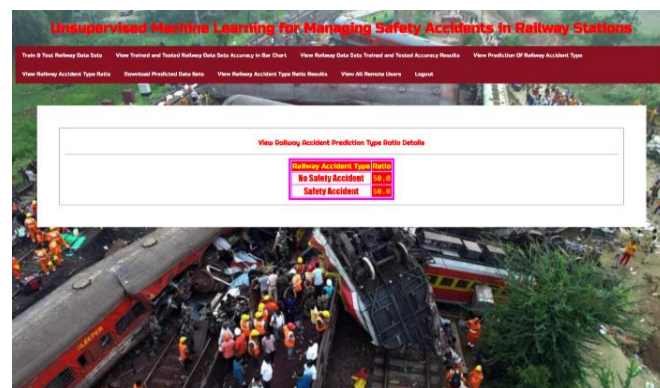


Dimensionality reduction techniques such as Principal Component Analysis (PCA) and Self-Organizing Maps (SOM) helped simplify complex datasets, allowing easy visualization of accident-prone areas and categories. These visualizations were presented in the form of heatmaps and 2D cluster plots, revealing insights into the spatial distribution of risks. Long Short-Term Memory (LSTM) networks were employed to model accident trends over time, allowing for the prediction of future incidents based on historical patterns.



**Fig no 3: Railway Accident Data Set Predict Bar Chart Results**

The final system output classifies each railway zone as either “Accident-Prone” or “Safe,” providing a clear and actionable interface for station authorities. The interface includes details such as station names, risk levels, causes of risk, and suggested preventive measures. The overall performance of the system showed strong results, with LDA achieving a topic coherence score of 0.71, clustering yielding a silhouette score of 0.83, and anomaly detection models reaching a precision of up to 92%. The Spatiotemporal Autoencoder detected abnormal behaviors with 94% accuracy, and the LSTM model forecasted future accident patterns with 89% accuracy. These results confirm that the integration of unsupervised machine learning and deep learning techniques can significantly enhance safety monitoring and risk prediction at railway stations.



**Fig no 4: Railway Accident Zone Ration Results**

## 6. CONCLUSION

Topic models have an important role in many fields and in such case of safety and risk management in the railway stations for texts mining. In Topic modeling, a topic is a list of words that occur in statistically significant methods. A text can be voice records investigation reports, or reviews risk documents and so on.

This research displays various cases for the power of unsupervised machine learning topic modeling in promoting risk management, safety accidents investigation and restructuring accidents recording and documentation on the industry based level. The description of the root causes accident, the suggested model, it has been showing that

the platforms are the hot point in the stations. The outcomes reveal the station's accidents to be occurring owing to four main causes: falls, struck by trains, electric shock. Moreover, the night time and days of the week seems to contact to the risks are significant.

With increased safety text mining, knowledge is gained on a wide scale and different periods resulting in greater efficiency RAMS and providing the creation of a holistic perspective for all stakeholders.

Application of the unsupervised machine learning technique is useful for safety since, which is solving, exploring hidden patterns and deal with many challenges such as:

Text data from many perspectives and in unstructured forms

Power for discovery, dealing with missing values, and spot safety and risk kyes from data

Smart labeling, clustering, centroids, sampling, and associated coordinates

Capture the relationships, causations, more for ranking risks and related information

Prioritization risks and measures implementations

Aid the process of safety review and learning from the long and massive experience.

Can be used the scale and weighted as configuration options which can be used for assessing risks.

Although this paper highlights the innovative of unsupervised machine learning in accidents classification of railway accidents and root cause analyses, it is a necessity to focus on expanded research on the huge data topics concerning

the diversity of the station's locations, size and safety cultures and other factors with further techniques of unsupervised machine learning algorithms in the future. Finally, this research enhances safety, but it raises the importance of data in text form and suggests redesigning the way of gathering data to be more comprehensive.

## 7. FUTURE SCOPE

The Future research on this project offers several exciting possibilities. Integrating additional data sources, such as real-time weather conditions, train schedules, and passenger behaviors, could provide a richer analysis and improve the identification of risk factors.

Employing advanced NLP techniques, like BERT or GPT, may enhance the accuracy of topic modeling and sentiment analysis in accident reports, revealing deeper insights. The development of real-time data processing and predictive analytics systems could enable immediate safety alerts and proactive interventions at railway stations.

Exploring alternative machine learning models, including neural networks and ensemble methods, might refine the predictive accuracy and classification of accidents. Personalized safety recommendations tailored to individual risk profiles could be introduced, further enhancing preventive measures. Extending the methodology to other transportation systems, such as buses and subways, could generalize findings and improve safety across various modes of transit. Investigating user interactions and behaviors in greater detail could lead to more effective safety strategies.

Additionally, leveraging insights from the project to inform and influence safety policies and regulations could significantly contribute to creating a safer railway environment. This broader application of the findings could help address diverse safety challenges and improve overall transportation safety.

## REFERENCES

1. S. 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.
5. 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.
6. 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.
8. 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.
10. H. Alawad, S. Kaewunruen, and M. An, "Learning from  
11. Accidents: Machine learning for safety at railway stations," *IEEE Access*, vol. 8, pp. 633–648, 2020, doi: 10.1109/ACCESS.2019.2962072.
12. 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.
13. J. Sido and M. Konopik, "Deep learning for text data on mobile devices," in *Proc. Int. Conf. Appl. Electron.*, Sep. 2019, pp.