RailVibes: A Novel Dataset of Rail Track Vibration for Derailment Detection

Adnan Quaium[1,2,*], Md. Abu Obaida Zishan[3,4], Debojit Pandit[1], Wasif Jalal[1], Md Roqunuzzaman Sojib[1], Shattik Islam Rhythm[1], Mashroor Hasan Bhuiyan[1], Jannatun Noor[1,3], and A. B. M. Alim Al Islam[1]

[1]Next-Generation Computing (NeC) Research Group, Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[2]Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh
[3]Computing for Sustainability and Social Good (C2SG) Research Group, Department of Computer Science and Engineering, School of Data and Sciences, United International University, Dhaka, Bangladesh
[4]Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
[*]Corresponding Author: Adnan Quaium (adnan.eee@aust.edu)

ABSTRACT

Railway transportation is crucial for socio-economic development, but derailments pose significant safety and economic risks. While existing research explores various derailment detection techniques, the lack of comprehensive datasets has hindered the utilization of track vibration data. This research presents a novel dataset comprising vibration data from 59 trains, including both freight and passenger trains with varying speeds and compositions. Additionally, the dataset includes data from 20 scenarios involving non-train sources, reflecting real-world operating conditions. Organized into 96 CSV files, the dataset contains 1.84 million sensor readings across eight channels. This diverse dataset, encompassing 768 train and non-train data samples, comprehensively represents the vibrational environment surrounding the railway track. To the best of our knowledge, this is the first dataset of this nature. It provides a valuable resource for the research community to develop and evaluate advanced machine learning algorithms for real-time derailment prediction. We believe this dataset will significantly contribute to developing robust and reliable derailment detection systems, ultimately enhancing railway safety and operational efficiency.

Background and Summary

Railway transportation plays a vital role in the socio-economic development of many countries[1]. However, railway accidents, including derailments[2], pose significant challenges, resulting in substantial economic losses, infrastructure damage, and, tragically, loss of human life. Derailments pose a significant threat to the safety and efficiency of railway operations, resulting in substantial economic losses, infrastructure damage, and, tragically, loss of human life. In developing countries (such as Bangladesh[3,4], India[5,6], Pakistan[7,8], Sri Lanka[9,10], several African countries[11–15], etc), where railway infrastructure plays a crucial role in transportation and economic development, the impact of derailments can be particularly severe. Early detection of potential derailment risks is paramount to ensure railway networks' safe and reliable operation. Early detection of potential derailment risks is crucial for ensuring the safety and reliability of railway operations.

Traditional methods for derailment detection, such as visual inspections[16–20] and track geometry measurements[21–26], are often time-consuming, labor-intensive, and may not provide real-time insights into track conditions. The existing research on derailment detection has primarily focused on techniques such as image processing[17,27–32], acoustic emission monitoring[33–38], etc. Advancements in sensor technologies and machine learning have enabled the development of novel approaches, such as tracking vibration data for real-time condition monitoring[39–44]. However, progress in this area has been hindered by the lack of large-scale, high-quality datasets representative of real-world operating conditions.

Besides, the dataset compiled by Chi Liu et al.[45] investigates the influence of train speed and track geometry on passenger ride comfort. By analyzing the relationship between train operating speeds and resulting vibrations, this dataset provides valuable insights into factors affecting passenger experience. However, its primary focus on ride comfort and passenger experience limits its applicability to derailment detection. Similarly, the dataset presented by Li Li et al.[46] concentrates on assessing the impact of vibrations on noise levels within railway vehicles. While this dataset offers valuable information regarding noise propagation and mitigation strategies, its focus on noise reduction precludes its direct application to derailment detection. On the other hand the dataset prepared by CR Soto-Ocampo et al.[47] focuses on the analysis of induced failures in rolling element bearings. This dataset includes vibration records collected from a test bench experiment, encompassing both normal operating conditions and conditions simulating rolling element defects. While valuable for research on bearing fault diagnosis, this dataset is not directly applicable to the detection of rail track faults associated with potential derailments.

In addition, the dataset developed by Yan Yan et al.[48] is specifically designed for rockfall detection and early warning systems along railway lines. This dataset leverages the characteristics of vibration signals to identify potential rockfall events. However, the focus on rockfall detection limits its applicability to the specific domain of derailment detection. These research works illustrate the domain-specific nature of existing datasets. While valuable for their respective research areas, they do not encompass the diverse range of vibrational phenomena associated with rail track faults and derailments. Therefore the utilization of track vibration data for derailment prediction has been relatively limited[49].

To address this critical research gap, we present a novel dataset specifically designed for derailment detection based on track vibration analysis. This dataset encompasses a wide range of scenarios, including vibration data collected from various train types (freight and passenger) with varying speeds and compositions. Moreover, we have incorporated data from diverse non-train sources, such as vibrations induced by vehicular crossings, pedestrian movements, adjacent track trains, and maintenance activities, which are particularly prevalent in developing countries. To the best of our knowledge, this is the first dataset of this nature, providing a valuable resource for the research community to develop and evaluate advanced derailment detection algorithms.

Our research is motivated by the urgent need for comprehensive and reliable datasets to drive innovation in derailment detection systems. By providing a unique and diverse dataset of track vibration data, we aim to facilitate the development and evaluation of advanced machine learning algorithms for real-time derailment prediction. In the process of this research, we make the following set of contributions in this paper.

- We developed a novel and diverse dataset for derailment detection based on track vibration analysis, encompassing real-world operating conditions, various train types (freight and passenger), speeds, and compositions.
- The dataset includes scenarios involving vibrations from non-train sources, such as vehicular crossings, pedestrian movements, adjacent track trains, and maintenance activities.
- Our dataset captures a broad range of environmental factors and operational disturbances to reflect realistic conditions.

Our dataset offers a valuable resource for the research community, supporting the development of more robust and reliable systems to improve railway safety and operational efficiency. Coupled with our proposed data collection and preprocessing methodology, it establishes a solid foundation for future research in this field. We believe this work will play a significant role in advancing derailment detection systems, ultimately contributing to safer and more efficient railway operations.

Methods

To accurately detect potential derailments, we focused on capturing and analyzing the vibrational signatures induced by approaching trains on the railway track. Under optimal track conditions, these vibrations exhibit predictable patterns that serve as a reference point for normal track behavior. However, any irregularities, such as defects or disruptions, can significantly alter these patterns, potentially leading to derailment.

To effectively classify track conditions, we aimed to collect a minimal yet comprehensive dataset of vibration data, so that it is straightforward to identify anomalies indicative of potential issues, including missing or damaged rail sections. To achieve this objective, we developed a specialized data collection module. This static device is strategically deployed on the railway track to capture high-resolution vibration data from passing trains. Equipped with sensitive sensors, the module records crucial information about the vibrational patterns and sends it to the data storage unit. The storage unit communicates wirelessly with the data collection unit. Figure 1 illustrates our proposed system for collecting railway track data.

To enhance the robustness of our dataset, we expanded our data collection to include a variety of non-train-related vibrational sources. This approach allows us to establish a robust baseline and effectively differentiate between normal train vibrations and anomalous patterns associated with potential derailments. By augmenting the dataset with vibrational data from diverse sources, such as side-track trains, maintenance activities, and vehicular crossings, the machine learning models can be trained to more accurately discriminate between normal and anomalous track conditions. This expanded dataset enables the models to develop a more robust understanding of the complex vibrational patterns associated with various operational scenarios, thereby improving their ability to detect potential derailment risks.

We focus on collecting vibration data along a 2000-meter stretch of railway track, enabling us to capture a comprehensive range of vibrational patterns associated with approaching trains over a significant distance. This extensive dataset provides valuable insights into the evolution of vibrational signatures as trains traverse the track.

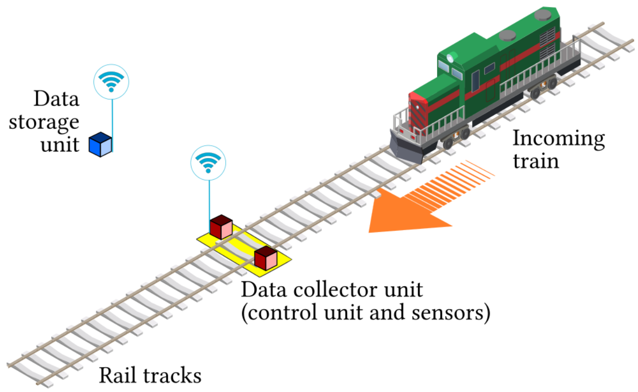
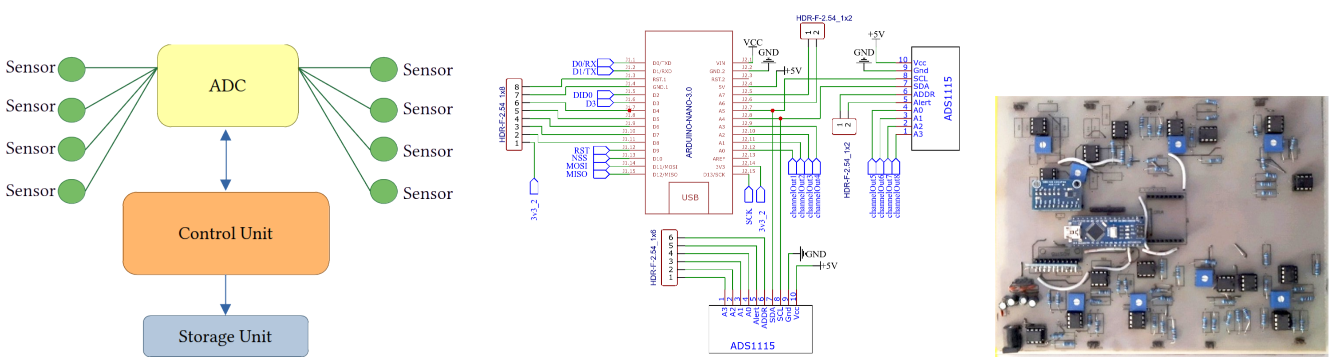


Figure 1. Conceptual diagram of the proposed data collection method

To ensure data quality and reliability, we conducted a rigorous data-cleaning process to identify and rectify errors and anomalies. This involved meticulous examination of the dataset to eliminate inconsistencies and missing values. Subsequently, we employed traditional machine learning algorithms to evaluate the efficacy of the refined dataset. By applying these established techniques, we aimed to assess the dataset's suitability for training and testing advanced machine learning models for accurate derailment detection.

Hardware Design

We design a data collection module comprising a microcontroller, piezoelectric sensors, an analog-to-digital converter (ADC), and a LoRa module. The module is the control unit, that is strategically positioned on the railway track to capture sensor data. We select piezoelectric sensors for their sensitivity in converting mechanical vibrations into electrical signals. The ATmega328 microcontroller, known for its low power consumption and versatility, is chosen to process the sensor data. To facilitate long-range communication between the control unit and the storage unit, we incorporate the SX1278 transceiver. The ADS1115 ADC ensures a high-resolution conversion of analog sensor signals into digital format. A 5-volt regulated power bank provides the necessary power supply for the control unit's operation. Figure 2 depicts the final configuration of the data collection module employed in our experiment. Figure 2a shows that there are a total of eight sensors used in our experiment. We deployed four piezoelectric sensors per rail track, resulting in eight sensors. This redundant sensor configuration ensures that the vibration data is not solely reliant on a single sensor. By employing multiple sensors, we enhance the system's robustness against noise and sensor failures. We determined that increasing the number of sensors beyond eight does not significantly improve data quality while unnecessarily expanding the device's cost and complexity.



(a) Block diagram of data collector module **(b)** Schematic diagram of the control unit **(c)** The final data collector module

Figure 2. The data collector module

Data Collection Procedure

We deliberately placed sensor nodes on the railway tracks at Banani Railway Station in Dhaka for our experimental setup. Figure 3 visually represents the experimental zone, with the red colored straight line indicating the railway tracks. The two tiny circular points on the map delineate the boundaries of our experimental area, spanning approximately 2000 meters.

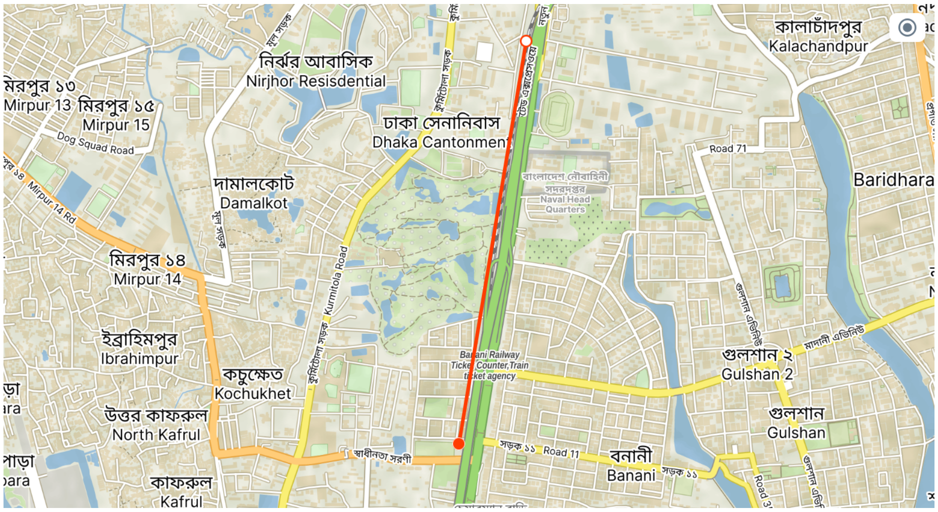


Figure 3. Place of experiment at Banani railway station, Dhaka, Bangladesh

As we discussed earlier, we deployed eight sensors along the railway track to capture comprehensive vibration data. These sensors were strategically positioned to ensure optimal coverage of the experimental area. By employing multi-sensor fusion, we integrated the data from these sensors to extract valuable insights into vibration patterns and potential anomalies. This technique offers several advantages, including improved accuracy, enhanced reliability, and possible cost reduction. By combining data from multiple sensors, we mitigate the impact of noise and individual sensor errors, leading to more precise and reliable detection of rail block conditions. Additionally, multi-sensor fusion enhances system reliability by reducing the impact of sensor failures. Although individual sensors may have limitations, multi-sensor fusion allows us to leverage lower-cost sensors, such as piezoelectric sensors, while achieving high performance.

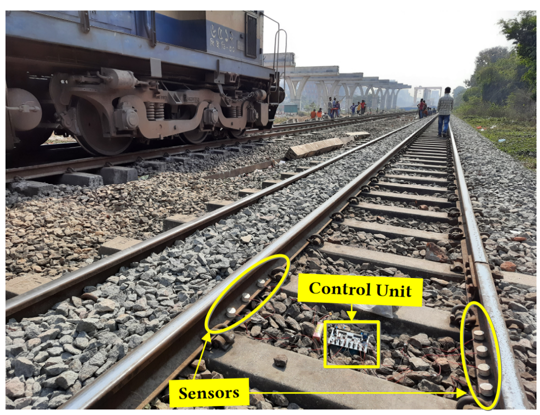


Figure 4. The data collector module deployed on the rail tracks

Figure 4 illustrates our real-world implementation at Banani Railway Station, showcasing the deployment of the data collection module, which incorporates piezoelectric sensor nodes strategically positioned along the railway tracks. This figure serves as a visual aid for comprehending the experimental setup and the spatial arrangement of the sensor nodes. As depicted in Figure 4, our experiment captured various conditions, including a passing train on an adjacent track and ongoing maintenance work. These factors contribute to the richness of our dataset, enhancing its robustness and generalizability. The presence of a nearby train introduces additional vibrational disturbances, while the maintenance activities simulate real-world scenarios where track conditions may deviate from the norm. This is one of several diverse scenarios we captured in our dataset. By incorporating these diverse conditions, our dataset becomes more representative of the complexities encountered in real-world railway operations.

Error Correction

We employ a novel framework for detecting and correcting errors within the dataset. This framework leverages a trained model to predict missing or erroneous data points. Specifically, the trained model is presented with the complete dataset and subsequently tasked with predicting the value of each data point individually, treating it as a missing entry. Using a defined evaluation metric, the predicted value is then compared to the ground truth value. If the model's prediction is deemed more suitable based on this evaluation, it replaces the ground truth value, effectively correcting the erroneous data point. This iterative process results in an error-corrected dataset.

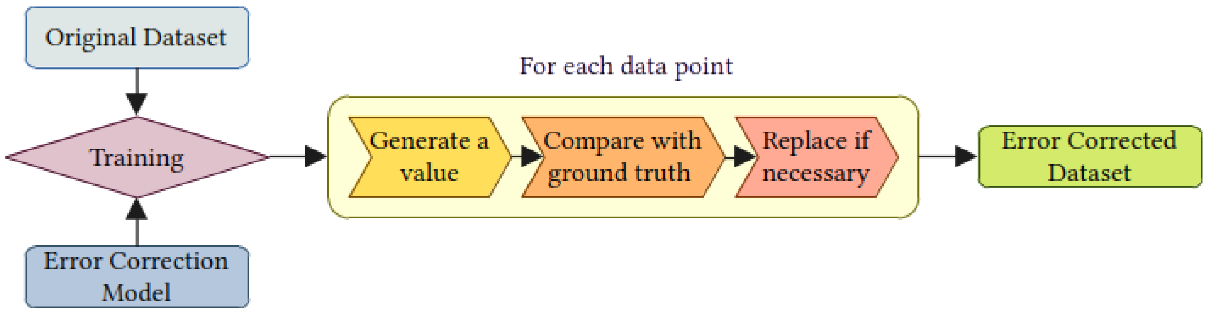


Figure 5. Error correction process

Figure 5 provides a visual representation of this error correction process. In this study, we utilize three distinct models for error correction: Long Short-Term Memory (LSTM)[50], Bidirectional Long Short-Term Memory (BiLSTM)[51], and Auto-Regressive Integrated Moving Average (ARIMA)[52]. LSTM and BiLSTM are prominent deep learning architectures, while ARIMA is a well-established statistical time series forecasting model. These models have been widely employed for missing value imputation tasks in various domains[53]. Notably, we exclusively utilize these models for error correction purposes within this research.

LSTM networks excel at capturing long-range temporal dependencies in time series data by employing gated mechanisms to regulate the flow of information. BiLSTM enhances this capability by processing the data in both forward and backward directions, enabling the model to capture contextual information from both past and future observations. ARIMA, on the other hand, leverages a combination of differencing, autoregression, and moving averages to effectively model and forecast stationary time series data.

Experimental Design

To assess the technical validity of our dataset, we employ a suite of classical machine learning algorithms, including Support Vector Classifier (SVC)[54], Logistic Regression (LR)[55], Decision Trees (Tree)[56], Random Forest (RF)[57], and K-Nearest Neighbors (KNN)[58]. These algorithms serve as a benchmark for evaluating the dataset's suitability for training and testing more sophisticated machine-learning models for rail condition monitoring. Our experimental protocol focuses on classifying vibration data into two distinct categories: "train-on-tracks" and "no-train-on-tracks." This binary classification task provides a fundamental assessment of the dataset's ability to discriminate between vibrations induced by train traffic and those originating from other sources, such as environmental disturbances or maintenance activities. We rigorously evaluate the performance of each model using standard metrics such as accuracy, precision, recall, and F1-score, providing insights into the dataset's effectiveness in supporting accurate and reliable classification of rail track conditions.

Data Records

The preparation of this dataset involved a rigorous data collection and curation process spanning over a year. Numerous challenges were encountered during data acquisition, including sensor malfunctions, partial data loss due to sensor failures, device damage, and instances of data disruption caused by external factors such as pedestrian interference. After meticulous data cleaning and quality control measures, the final dataset comprises vibration data collected from 59 approaching trains, encompassing both freight and passenger trains with varying speeds and compositions. Furthermore, it includes data from 20 scenarios involving non-train sources, such as vehicular crossings, pedestrian movements, adjacent track trains, and maintenance activities, reflecting real-world operating conditions. This diverse dataset is organized into 96 CSV files, each containing eight channels of time-series vibration data with approximately 2400 data points per channel, resulting in a substantial collection of approximately 1.84 million sensor readings. These diverse data sources, encompassing 768 samples of train and non-train data, provide a comprehensive representation of the complex vibrational environment surrounding the railway track.

Data Cleansing

We collected a substantial dataset of vibration data associated with approaching trains. Upon thorough examination, we identified instances of incomplete data, primarily attributed to sensor malfunctions or data acquisition failures. These instances, where data from one or more of the eight sensors was missing, were excluded from the dataset to ensure data integrity. After a rigorous quality control process, the final dataset comprised vibration data from 59 train approaches, each with complete readings from all eight sensors. To enrich the dataset and enhance the model's ability to differentiate between train-induced vibrations and other sources of ground motion, we also included data collected under various conditions, such as when no train was present, during pedestrian crossings, and when trains passed on adjacent tracks. This diverse dataset provides a comprehensive representation of the vibrational environment surrounding the railway track. The final dataset comprises vibration data from 59 approaching trains, each with complete readings from all eight sensors.

Details and Types of Data

The database comprises multiple CSV files, each containing eight columns corresponding to the vibration data collected by the eight piezoelectric sensors. Piezoelectric sensors function based on the piezoelectric effect, where mechanical stress, such as pressure or force, induces an electrical charge within the sensor material. This phenomenon arises from the deformation of the piezoelectric material under applied stress, resulting in the generation of an electrical charge proportional to the applied force. This electrical charge is subsequently measured as a voltage signal, allowing for the detection of variations in force, pressure, acceleration, or strain. Consequently, each column in the CSV file records the vibration data as a voltage signal, specifically in millivolts.

Sl_No	Sensor_1	Sensor_2	Sensor_3	Sensor_4	Sensor_5	Sensor_6	Sensor_7	Sensor_8
0	32	24	44	24	40	56	40	48
1	32	24	44	24	44	56	40	40
2	32	28	44	32	40	52	40	48
3	32	24	44	12	48	56	40	40
4	32	24	44	40	40	56	36	48
5	32	28	44	40	48	56	40	40
6	28	28	44	12	48	56	40	48
7	32	24	44	36	40	60	40	44
8	32	24	44	36	40	56	40	52
9	32	24	44	36	52	56	36	44
10	32	28	44	24	40	60	40	52
				... ...				
2442	778	782	782	782	0	782	0	0
2443	0	782	0	782	0	782	778	0
2444	0	778	778	0	782	782	32	778
2445	778	778	0	0	0	0	52	778
2446	778	782	782	0	782	0	60	0
2447	100	0	8	0	782	782	36	0
2448	0	782	782	782	782	8	56	778
2449	774	0	778	778	774	778	68	0
2450	778	0	0	782	782	782	4	778
2451	778	782	726	0	782	782	24	774
2452	778	782	782	782	782	782	12	778
2453	778	778	16	782	782	778	36	770

Figure 6. Structure of a single CSV file

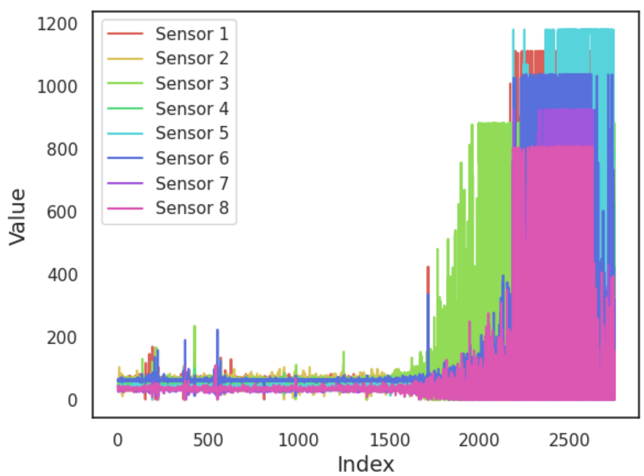
The general structure of each CSV file in the dataset is shown in Figure 6. *Sl_No* column shows the serial number of the data points. The rest of the columns ($Sensor\_1, Sensor\_2, Sensor\_3, Sensor\_4, Sensor\_5, Sensor\_6, Sensor\_7, Sensor\_8$) denote the vibration data from an approaching train. The first row in each file represents the vibration data captured when the train is 2000 meters away from the sensor array, while the final row corresponds to the moment when the train directly passes over the sensors. We categorized the collected data into two distinct subsets:

- Train data: This subset encompasses vibration data generated by approaching trains.
- Non-train data: This subset includes vibration data originating from sources other than approaching trains, such as vibrations caused by vehicles crossing the track, trains passing on adjacent tracks, and other environmental disturbances.

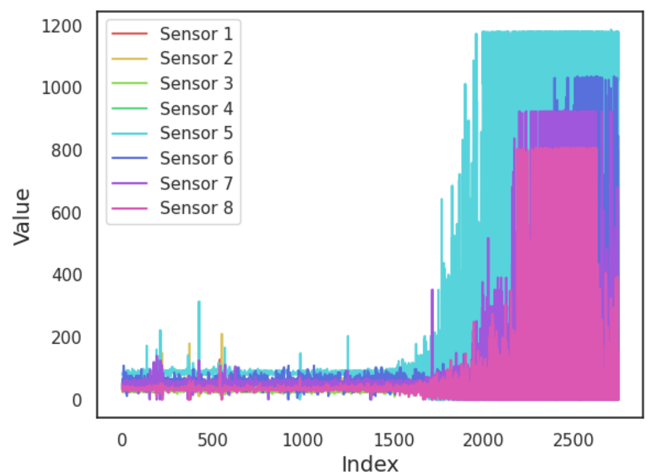
The final dataset consists of 59 CSV files containing vibration data from approaching trains (each file represents vibration data from a train) and 20 files containing vibration data from non-train sources.

Data Diversity

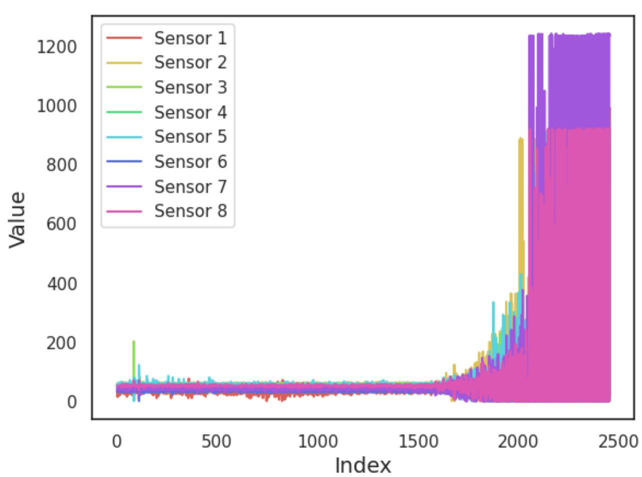
Our research utilizes a comprehensive dataset of vibration data collected from 60 train approaches. After rigorous data preprocessing, the final dataset comprises 59 train approaches. This dataset exhibits significant diversity, encompassing both freight and passenger trains. The freight trains, traveling at an average speed of 25 kmph (approximate), exhibited a range of compositions, varying from 28 to 38 compartments. In contrast, the passenger trains, operating at an average speed of 73 kmph (approximate), consisted of 8 to 17 compartments. Figure 7 shows the visual representation of the sample vibration data for two freight trains and two passenger trains, demonstrating variations in signal patterns across different train types.



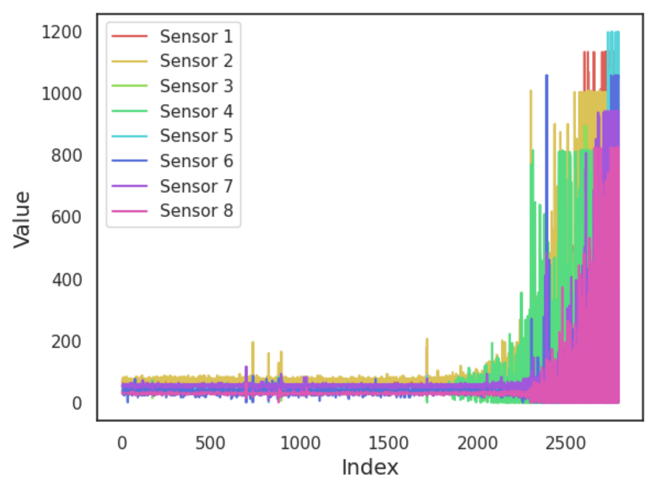
(a) Data from an approaching freight train



(b) Data from another approaching freight train



(c) Data from an approaching passenger train



(d) Data from another approaching passenger train

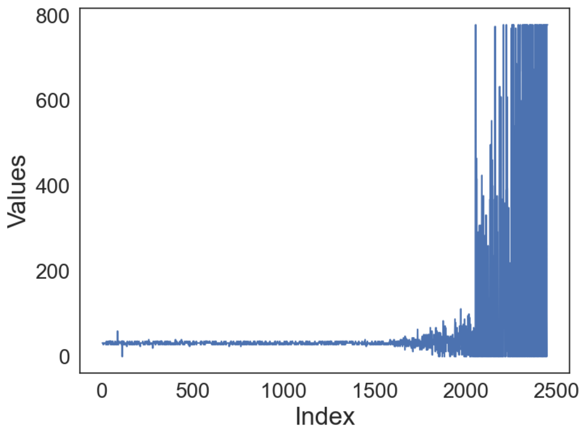
Figure 7. Sample vibration data recorded by eight sensors for two freight trains and two passenger trains

Beyond train-induced vibrations, the dataset incorporates a diverse array of non-train sources. These include vibrations generated by pedestrians, maintenance activities, adjacent track trains, and vehicles traversing the railway crossing, etc. Notably, the dataset also includes vibration data collected from an uprooted rail section situated adjacent to the operational track, providing valuable insights into the vibrational signatures associated with track abnormalities.

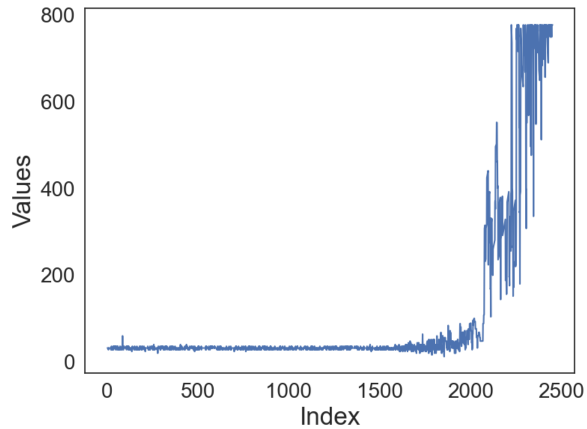
By encompassing a wide spectrum of train types, operational conditions, and environmental disturbances, this dataset provides a comprehensive and valuable resource for the development and evaluation of robust rail condition monitoring systems.

Faulty Data

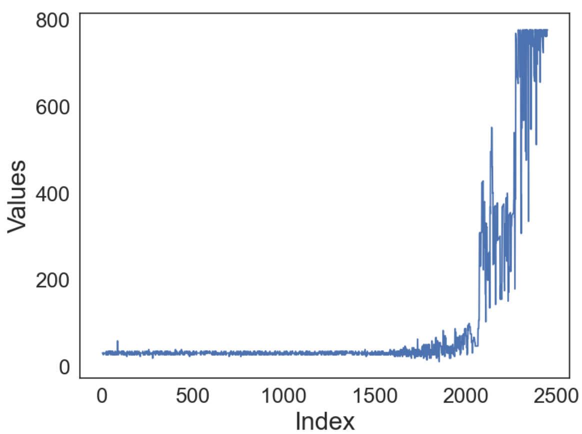
Our dataset exhibited instances of missing values and potential errors. To enhance data quality, we implemented data imputation and correction techniques as detailed in the "Error Correction" subsection.



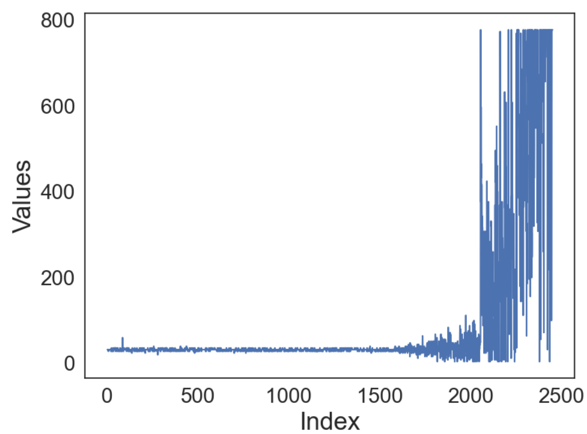
(a) Raw Data



(b) Corrected using LSTM



(c) Corrected using BiLSTM



(d) Corrected using ARIMA

Figure 8. Comparison between raw data and corrected data

Figure 8 presents a comparative analysis of raw and corrected sensor data for a single train, utilizing three imputation methods: LSTM, BiLSTM, and ARIMA. Figure 8a illustrates the raw sensor data, where missing values are represented by zero, resulting in a sparser data representation. Figures 8b, 8c, and 8d depict the data after imputation and correction using the respective models. These corrected datasets demonstrate a smoother and more consistent representation of the underlying vibration patterns compared to the raw data.

Table 1. Comparison of data point corrections by LSTM, BiLSTM, and ARIMA

Model Name	Imputed Points	Corrected Points	Total Modified Points	Total Data Points
LSTM	112311 (8.9%)	96595 (7.65%)	208906 (16.55%)	1262448 (100%)
BiLSTM		94042 (7.5%)	206353 (16.4%)	
ARIMA		60274 (4.77%)	172585 (13.67%)	

Table 1 summarizes the number of data points modified by each imputation method. The original dataset comprised 1,262,448 data points, with approximately 8.9% exhibiting missing values. All three models successfully imputed these missing values. Furthermore, the LSTM model corrected 96,595 data points (7.65% of the total), the BiLSTM model corrected 94,042 data points (7.5%), and the ARIMA model corrected 60,274 data points (4.77%).

Technical Validation

Technical validation with traditional machine learning models serves several critical purposes. Firstly, it provides a rigorous assessment of data quality by identifying potential biases, anomalies, and inconsistencies. Secondly, it establishes a baseline for comparison with more complex models, enabling the evaluation of performance gains. Thirdly, it facilitates hyperparameter tuning and feature selection by providing insights into model behavior and feature importance. Finally, the interpretability of some traditional models offers valuable insights into the underlying data patterns. Collectively, these factors underscore the significance of utilizing traditional machine learning models for comprehensive dataset validation, ultimately contributing to the development of a more robust and reliable dataset. For model training, validation, and evaluation, we employed a suite of five machine learning models implemented within the scikit-learn library[59], such as Support Vector Classifier (SVC), Logistic Regression (LR), Decision Trees (Tree), Random Forest (RF), and K-Nearest Neighbors (KNN).

For SVC and LR, we explored both convex optimization and gradient-descent-based optimization techniques[60–62]. The gradient-descent optimization utilizes the stochastic gradient descent (SGD) algorithm, configured with a maximum iteration limit of $10,000$. Both convex and gradient-descent-based methods use a tolerance of $1e^{-5}$. For K-Nearest Neighbors, we employ an Euclidean distance metric and utilized three nearest neighbors for classification. For the Decision Tree and Random Forest models, we utilize the default hyperparameter settings provided by the scikit-learn library.

Preprocessing

To ensure optimal model training and convergence, we implemented several preprocessing steps. The dataset comprises 59 CSV files containing vibration data from approaching trains and 20 CSV files containing non-train-data data (no approaching trains). Each file contains time-series data recorded by an array of eight sensors.

To guarantee a sufficient stopping distance for approaching trains, we excluded the final 700 data points from each train data file. This exclusion, equivalent to approximately 600 meters, ensures that the system can effectively alert the train control system before the train reaches the sensor array.

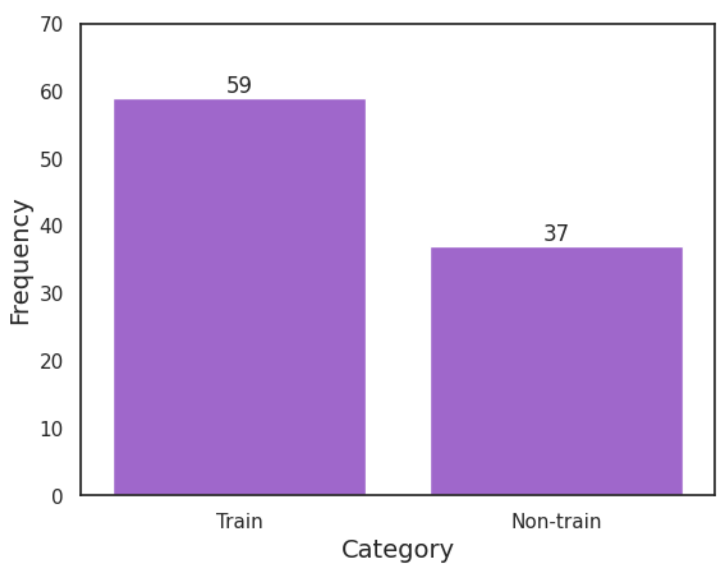


Figure 9. Data Distribution

To ensure data homogeneity across different train approaches, we standardized the length of the input sequences. Given the variability in train speeds, the original CSV files exhibited varying lengths. To address this, we extracted the last 1,754 samples from each train data file as per the length of the shortest incoming train file after excluding 700 samples. This approach ensures consistent input lengths for all train-data samples. Conversely, for the non-train-data, which is not time-sensitive in our context, we extracted non-overlapping windows of 1,754 samples per sensor from each non-train-data file, resulting in 59 and 37 samples for train-data (labeled 1) and non-train-data (labeled 0), respectively, as illustrated in Figure 9.

We applied feature engineering techniques to enhance the discriminability between train-data and non-train-data. As shown in Figure 10, the raw vibration data (data with no error correction) exhibits limited distinguishability between the two classes. To address this, we computed the time-series data's first-order derivative (difference) for each sensor channel and subsequently extracted the maximum value across all eight channels[63]. This feature engineering step significantly improved the separability between train-data and non-train-data data. Finally, we applied standard scaling (zero means, unit variance)[64,65] to the preprocessed data before model training and evaluation. A detailed overview of the entire preprocessing pipeline is presented in Algorithm 1.

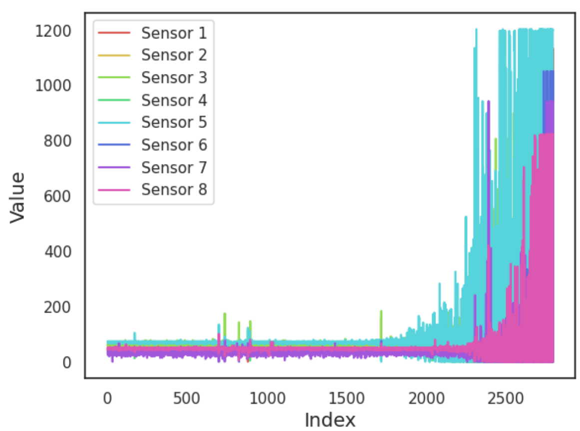
Algorithm 1 Data Preprocessing

Require: inputs: A 3D Array where:

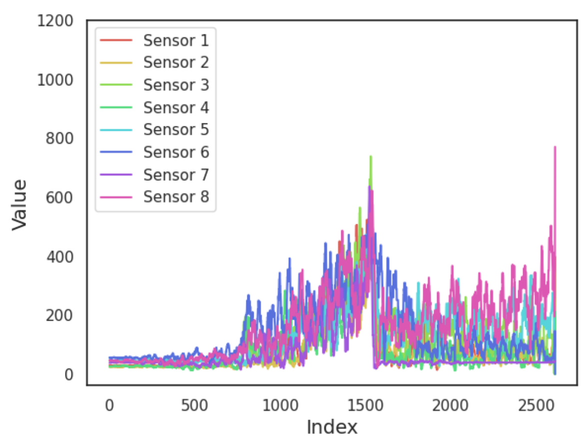
- First dimension: number of CSV files
- Second dimension: the length of the corresponding CSV file (number of rows).
- Third dimension: Number of sensors/channels (columns of the CSV File)

Ensure: inputs: A preprocessed 2D array for training our collection of machine-learning models

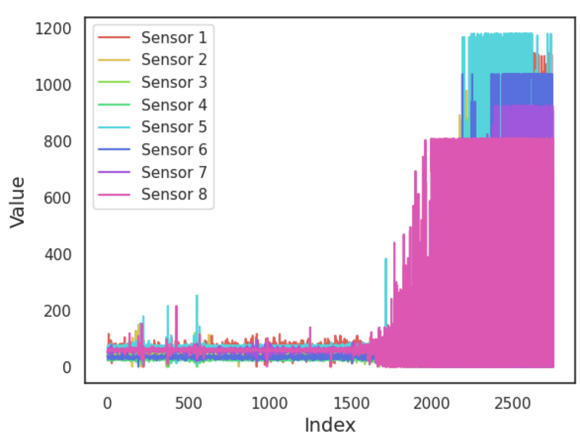
```
stopping_samples ← 700
minimum_file_length ← 2454
minimum_array_length ← minimum_file_length-stopping_samples      ▷ Equals 1754
inputs ← inputs[:, 0:-stopping_samples,:]      ▷ Removing the last 700 values from each file across all sensors
inputs ← inputs[:, -minimum_array_length:,:]      ▷ Selecting the last 1754 values from each file across all sensors
inputs ← inputs.diff(axis=1)      ▷ Computing the first-order derivative/difference along the second axis
inputs ← inputs.max(axis=-1)      ▷ Find the maximum along the last axis
inputs ← standard_scaling(inputs)      ▷ Apply standard scaling to normalize the data
```



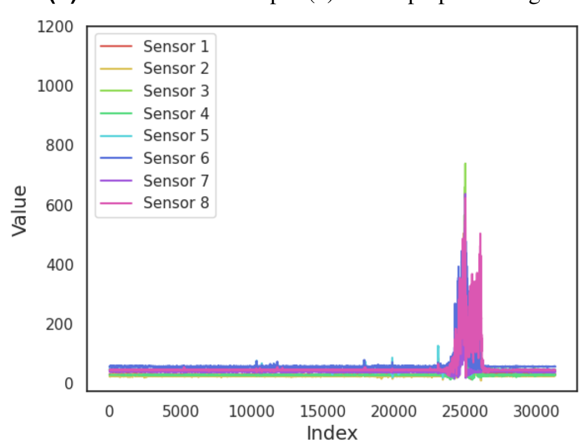
(a) Train-data sample-(1) before preprocessing



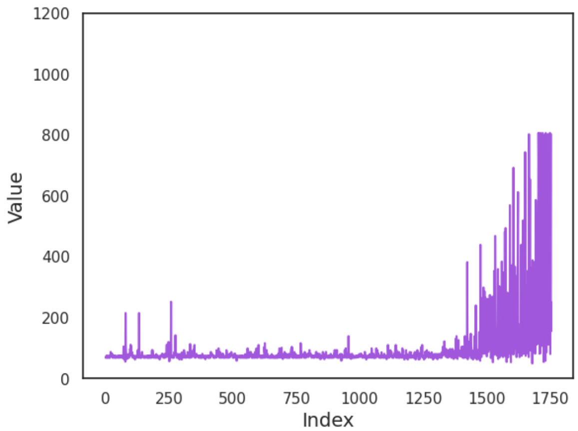
(b) Non-train-data sample-(1) before preprocessing



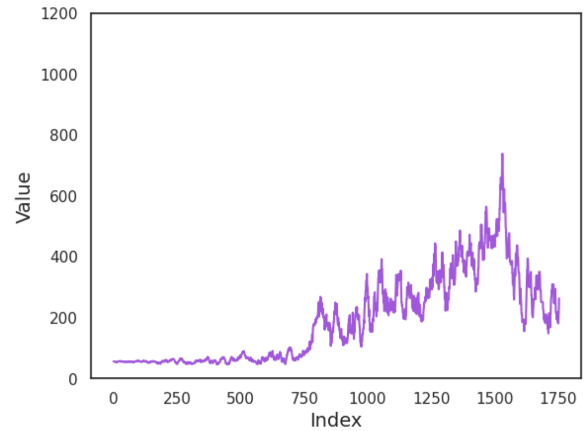
(c) Train-data sample-(2) before preprocessing



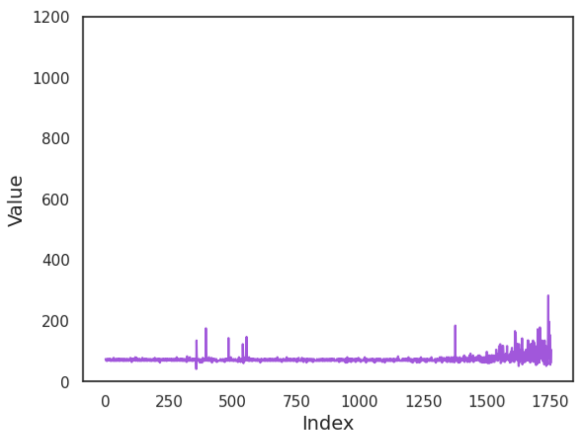
(d) Non-train-data sample-(2) before preprocessing



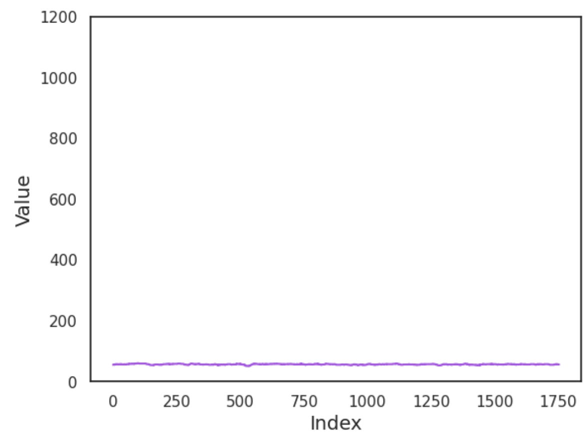
(e) Train-data sample-(1) after preprocessing



(f) Non-train-data sample-(1) after preprocessing



(g) Train-data sample-(2) after preprocessing



(h) Non-train-data sample-(2) after preprocessing

Figure 10. Data visualization before (a,b,c,d) and after (e,f,g,h) preprocessing through Algorithm 1

Table 2. Performance on raw data *before* preprocessing

Models	Precision	Recall	F1-Score	Accuracy
SVC-SGD	0.856	0.864	0.85	0.854
SVC-Convex	1.0	1.0	1.0	1.0
LR-SGD	0.898	0.897	0.88	0.884
LR-Convex	1.0	1.0	1.0	1.0
Decision-Tree	0.927	0.927	0.923	0.926
RF	0.945	0.932	0.932	0.937
KNN	0.89	0.893	0.887	0.895

Table 3. Performance on raw data *after* preprocessing

Models	Precision	Recall	F1-Score	Accuracy
SVC-SGD	0.898	0.903	0.892	0.895
SVC-Convex	1.0	1.0	1.0	1.0
LR-SGD	0.924	0.930	0.924	0.927
LR-Convex	1.0	1.0	1.0	1.0
Decision-Tree	0.992	0.985	0.988	0.989
RF	0.984	0.973	0.977	0.979
KNN	0.984	0.973	0.977	0.979

Experimental Results

To rigorously evaluate our dataset, we employed a five-fold stratified cross-validation approach[66,67] to our dataset. This technique ensures that each fold maintains the original class distribution of the dataset, mitigating potential biases in the model evaluation. Model performance was assessed using macro-averaged metrics, including accuracy, F1-score, precision, and recall, computed across the five folds.

Tables 2 and 3 demonstrate the significant impact of our data preprocessing pipeline on model performance. Notably, the application of the first-order derivative (difference) followed by the selection of the maximum value across all sensor channels[63] resulted in a substantial improvement in classification performance across all evaluated models. This observation underscores the effectiveness of our feature engineering approach in enhancing the discriminative power of the vibration data.

Table 4. Performance on data corrected by ARIMA

Models	Precision	Recall	F1-Score	Accuracy
SVC-SGD	0.931	0.939	0.934	0.937
SVC-Convex	1.0	1.0	1.0	1.0
LR-SGD	0.932	0.939	0.934	0.937
LR-Convex	1.0	1.0	1.0	1.0
Decision-Tree	0.992	0.985	0.988	0.989
RF	0.984	0.973	0.977	0.979
KNN	0.984	0.973	0.977	0.979

Table 5. Performance on data corrected by LSTM

Models	Precision	Recall	F1-Score	Accuracy
SVC-SGD	0.931	0.939	0.934	0.937
SVC-Convex	1.0	1.0	1.0	1.0
LR-SGD	0.903	0.905	0.893	0.895
LR-Convex	1.0	1.0	1.0	1.0
Decision-Tree	0.992	0.985	0.988	0.989
RF	0.984	0.973	0.977	0.979
KNN	0.984	0.973	0.977	0.979

Table 6. Performance on data corrected by BiLSTM

Models	Precision	Recall	F1-Score	Accuracy
SVC-SGD	0.931	0.939	0.934	0.937
SVC-Convex	1.0	1.0	1.0	1.0
LR-SGD	0.903	0.905	0.893	0.895
LR-Convex	1.0	1.0	1.0	1.0
Decision-Tree	0.992	0.985	0.988	0.989
RF	0.984	0.973	0.977	0.979
KNN	0.984	0.973	0.977	0.979

Tables 4, 5, and 6 present the performance metrics of the trained models on datasets corrected using the ARIMA, LSTM, and BiLSTM imputation methods, respectively. Analysis of these results reveals several key insights. Firstly, we observe consistently high performance across all three tables for most models, particularly SVC-Convex, LR-Convex, Tree, and RF, indicating the overall high quality of the dataset and the suitability of the chosen machine learning algorithms for this classification task.

Secondly, the impact of data correction techniques on model performance is evident. While some models exhibit consistently high performance across all corrected datasets, others, such as LR-SGD, show variations in performance depending on the specific correction method. This observation suggests that the sensitivity of different models to data correction techniques may vary. Despite these variations, the overall system robustness appears to be maintained across different correction methods, highlighting the effectiveness of the data correction techniques in improving data quality and enhancing model performance.

The consistently high performance achieved by most models across all corrected datasets underscores the effectiveness of the data correction techniques (ARIMA, LSTM, and BiLSTM) in addressing data quality issues and improving the overall data quality. This analysis provides valuable insights into the impact of data correction techniques on model performance and the overall robustness of the proposed system. The results suggest that models such as SVC-Convex, LR-Convex, Tree, and RF are strong candidates for deployment in the final system due to their consistently high performance across different correction scenarios. Furthermore, the robustness of the system across different correction methods allows for flexibility in selecting the most suitable data correction technique based on factors such as computational cost, ease of implementation, and specific performance requirements.

Usage Notes

The dataset used and analyzed in this study is currently not publicly available. However, it can be obtained from the corresponding author upon reasonable request. This dataset contains vibration data collected from railway tracks in Banani, Dhaka, Bangladesh. It includes data from 59 train approaches, encompassing both freight and passenger trains with varying speeds and compositions. Additionally, the dataset incorporates vibration data from various non-train sources, such as vehicular crossings, pedestrian movements, adjacent track trains, and maintenance activities.

Code availability

A subset of the larger, comprehensive dataset of track vibration data collected for derailment detection research is available on GitHub[68] for exploratory analysis. The complete dataset used and analyzed during the current study can be obtained from the corresponding author upon reasonable request and justification of research needs.

References

1. Quium, A. A. Transport corridors for wider socio–economic development. *Sustainability* **11**, 5248 (2019).
2. Duncan, I. G., McCann, J. & Brown, A. The investigation of derailments. *Proc. Inst. Mech. Eng.* **191**, 323–332 (1977).
3. UNB News. Nationwide train delays as panchagarh express derailed in dhaka. *United News Bangladesh* (2024). Available at: https://unb.com.bd/category/Bangladesh/nationwide-train-delays-as-panchagarh-express-derailed-in-dhaka/145637 (Accessed: 14 January, 2025).
4. UNB. Comilla train derailment: 4 bogies removed. *Dhaka Tribune* (2024). Available at: https://www.dhakatribune.com/bangladesh/nation/342156/comilla-train-derailment-4-bogies-removed (Accessed: 14 January, 2025).
5. BSS. Two killed, 20 injured in india train derailment. *BSS News* (2024). Available at: https://www.bssnews.net/news-flash/201758 (Accessed: 14 January, 2025).

6. Mishra, P. Train derailment bids on rise, 18 attempts just since august: Railways. *India Today* (2024). Available at: https://www.indiatoday.in/india/story/indian-railways-train-derailment-attempts-2596944-2024-09-10 (Accessed: 14 January, 2025).

7. Walsh, A. Pakistan passenger train derails killing 30. *BBC News* (2023). Available at: https://www.bbc.com/news/world-asia-66421481 (Accessed: 14 January, 2025).

8. News Desk. Freight train derailment in kotri disrupts karachi-lahore rail traffic. *24 News* (2024). Available at: https://24newshd.tv/29-Aug-2024/freight-train-derailment-in-kotri-disrupts-karachi-lahore-rail-traffic (Accessed: 14 January, 2025).

9. Staff Writer. Train derails near maradana;railway operations delayed. *News 1st* (2024). Available at: https://www.newsfirst.lk/2024/10/15/train-derails-near-maradana-railway-operations-delayed (Accessed: 14 January, 2025).

10. IANS. Train derailment disrupts coastal line operations in sri lanka. *Lokmat Times* (2024). Available at: https://www.lokmattimes.com/international/train-derailment-disrupts-coastal-line-operations-in-sri-lanka-1 (Accessed: 14 January, 2025).

11. Ahmed, A. Nrc confirms warri-itakpe train derailment, assures service restoration. *News Cent.* (2024). Available at: https://newscentral.africa/nrc-confirms-warri-itakpe-train-derailment-assures-service-restoration/ (Accessed: 14 January, 2025).

12. Rédaction Africanews. Tunisia train derailment kills two, injures 34. *Afr. News* (2024). Available at: https://www.africanews.com/2023/06/21/tunisia-train-derailment-kills-two-injures-34/ (Accessed: 14 January, 2025).

13. Vaidyanathan, V. Train derailment in paarl, south africa leaves many injured: What we know so far. *Times Now World* (2024). Available at: https://www.timesnownews.com/world/train-derailment-in-paarl-south-africa-leaves-many-injured-what-we-know-so-far-article-113437101 (Accessed: 14 January, 2025).

14. Bensekkaim, K. Tanzania: 70 injured in train derailment. *Al 24 News* (2024). Available at: https://al24news.com/en/tanzania-70-injured-in-train-derailment/ (Accessed: 14 January, 2025).

15. Xinhua. 25 injured as train derails in south africa's western cape. *Dly. News Egypt* (2024). Available at: https://www.dailynewsegypt.com/2024/09/18/25-injured-as-train-derails-in-south-africas-western-cape/ (Accessed: 14 January, 2025).

16. Gibert, X., Patel, V. M. & Chellappa, R. Robust fastener detection for autonomous visual railway track inspection. In *2015 IEEE winter conference on applications of computer vision*, 694–701 (IEEE, 2015).

17. Kumar, A. & Harsha, S. A systematic literature review of defect detection in railways using machine vision-based inspection methods. *Int. J. Transp. Sci. Technol.* (2024).

18. Li, Y., Trinh, H., Haas, N., Otto, C. & Pankanti, S. Rail component detection, optimization, and assessment for automatic rail track inspection. *IEEE Transactions on Intell. Transp. Syst.* **15**, 760–770 (2013).

19. Steets, P. G. & Tse, Y. H. Conrail's integrated automated wayside inspection. In *Proceedings of the 1998 ASME/IEEE Joint Railroad Conference*, 113–125 (IEEE, 1998).

20. Babenko, P. *Visual inspection of railroad tracks* (University of Central Florida, 2009).

21. Chen, Q. *et al.* A railway track geometry measuring trolley system based on aided ins. *Sensors* **18**, 538 (2018).

22. Weston, P., Roberts, C., Yeo, G. & Stewart, E. Perspectives on railway track geometry condition monitoring from in-service railway vehicles. *Veh. system dynamics* **53**, 1063–1091 (2015).

23. Akpinar, B. & Gulal, E. Multisensor railway track geometry surveying system. *IEEE Transactions on Instrumentation Meas.* **61**, 190–197 (2011).

24. Guler, H., Jovanovic, S. & Evren, G. Modelling railway track geometry deterioration. In *Proceedings of the Institution of Civil Engineers-Transport*, vol. 164, 65–75 (Thomas Telford Ltd, 2011).

25. Higgins, C. & Liu, X. Modeling of track geometry degradation and decisions on safety and maintenance: A literature review and possible future research directions. *Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit* **232**, 1385–1397 (2018).

26. Gonzalo, A. P., Horridge, R., Steele, H., Stewart, E. & Entezami, M. Review of data analytics for condition monitoring of railway track geometry. *IEEE Transactions on Intell. Transp. Syst.* **23**, 22737–22754 (2022).

27. Jiang, Y., Mozumder, S. A., Ma, C. & Rob, M. A. Derailment detection of mining shaft's rail vehicle using machine vision on edge device. In *2024 3rd International Conference on Innovations and Development of Information Technologies and Robotics (IDITR)*, 27–31 (IEEE, 2024).

28. Schlake, B. W. *et al.* Machine vision condition monitoring of heavy-axle load railcar structural underframe components. *Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit* **224**, 499–511 (2010).

29. Ahuja, N. & Barkan, C. P. *Machine vision for railroad equipment undercarriage inspection using multi-spectral imaging* (IDEA Programs, Transportation Research Board, 2007).

30. Camargo, L. F. M. *et al.* Machine vision inspection of railroad track. Tech. Rep., NEXTRANS Center (US) (2011).

31. Hart, J. M., Ahuja, N., Barkan, C. P. & Davis, D. D. A machine vision system for monitoring railcar health: Preliminary results. *Technol. Dig. TD-04-008, Assoc. Am. Railr. Pueblo, Colo.* (2004).

32. Li, Y., Otto, C., Haas, N., Fujiki, Y. & Pankanti, S. Component-based track inspection using machine-vision technology. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 1–8 (2011).

33. Thakkar, N., Steel, J. & Reuben, R. Rail–wheel interaction monitoring using acoustic emission: A laboratory study of normal rolling signals with natural rail defects. *Mech. Syst. Signal Process.* **24**, 256–266 (2010).

34. Bruzelius, K. & Mba, D. An initial investigation on the potential applicability of acoustic emission to rail track fault detection. *Ndt & E Int.* **37**, 507–516 (2004).

35. Janeliukstis, R., Clark, A., Papaelias, M. & Kaewunruen, S. Flexural cracking-induced acoustic emission peak frequency shift in railway prestressed concrete sleepers. *Eng. Struct.* **178**, 493–505 (2019).

36. Diana, G. *et al.* Full-scale derailment tests on freight wagons. *Veh. system dynamics* **60**, 1849–1866 (2022).

37. Janeliukstis, R. & Kaewunruen, S. A novel separation technique of flexural loading-induced acoustic emission sources in railway prestressed concrete sleepers. *IEEE Access* **7**, 51426–51440 (2019).

38. Zhang, X. *et al.* An investigation on acoustic emission detection of rail crack in actual application by chaos theory with improved feature detection method. *J. Sound Vib.* **436**, 165–182 (2018).

39. Chakraborty, T., Khan, T. A. & Islam, A. A. A. Poster: Railcop: Detecting missing rail on railway using wireless sensor networks. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion*, MobiSys '16 Companion, 16, 10.1145/2938559.2948844 (Association for Computing Machinery, New York, NY, USA, 2016).

40. Chakraborty, T. *et al.* A New Network Paradigm for Low-Cost and Lightweight Real-Time Communication between Train and Rail Track to Detect Missing and Faulty Rail Blocks. *J. Netw. Comput. Appl.* **131**, 40 – 54, https://doi.org/10.1016/j.jnca.2019.01.018 (2019).

41. Nurain, N., Tairin, S., Khan, T. A., Ishraq, S. & Al Islam, A. A. Power Attack: An Imminent Security Threat in Real-Time System for Detecting Missing Rail Blocks in Developing Countries. *Comput. & Secur.* **84**, 35 – 52, https://doi.org/10.1016/j.cose.2019.03.012 (2019).

42. Salim, S. I. *et al.* Long-range low-cost networking for real-time monitoring of rail tracks in developing countries. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, ICTD '22, 10.1145/3572334.3581765 (Association for Computing Machinery, New York, NY, USA, 2023).

43. Celiński, I., Burdzik, R., Młyńczak, J. & Kłaczyński, M. Research on the applicability of vibration signals for real-time train and track condition monitoring. *Sensors* **22**, 2368 (2022).

44. Lin, C.-C. & Zhuang, Z.-Y. Qualifying data on railroad track vibrations: a hybrid data preprocessing flow of statistical and machine learning approaches. *J. Chin. Inst. Eng.* **46**, 839–855 (2023).

45. Liu, C., Thompson, D., Griffin, M. J. & Entezami, M. Effect of train speed and track geometry on the ride comfort in high-speed railways based on iso 2631-1. *Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit* **234**, 765–778 (2020).

46. Li, L. *et al.* Influence of rail fastener stiffness on railway vehicle interior noise. *Appl. Acoust.* **145**, 69–81 (2019).

47. Soto-Ocampo, C. R., Mera, J. M., Cano-Moreno, J. D. & Garcia-Bernardo, J. L. Low-cost, high-frequency, data acquisition system for condition monitoring of rotating machinery through vibration analysis-case study. *Sensors* **20**, 3493 (2020).

48. Yan, Y., Li, T., Liu, J., Wang, W. & Su, Q. Monitoring and early warning method for a rockfall along railways based on vibration signal characteristics. *Sci. reports* **9**, 6606 (2019).

49. Pappaterra, M. J., Flammini, F., Vittorini, V. & Bešinović, N. A systematic review of artificial intelligence public datasets for railway applications. *Infrastructures* **6**, 136 (2021).

50. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780, 10.1162/neco.1997.9.8.1735 (1997). https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.
51. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Process.* **45**, 2673–2681 (1997).
52. Box, G. & Jenkins, G. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing (Holden-Day, 1976).
53. Fang, C. & Wang, C. Time series data imputation: A survey on deep learning approaches. *arXiv preprint arXiv:2011.11347* (2020).
54. Xue, H., Yang, Q. & Chen, S. Svm: Support vector machines. In *The top ten algorithms in data mining*, 51–74 (Chapman and Hall/CRC, 2009).
55. Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M. & Klein, M. *Logistic regression* (Springer, 2002).
56. Rokach, L. & Maimon, O. Decision trees. *Data mining knowledge discovery handbook* 165–192 (2005).
57. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).
58. Kramer, O. & Kramer, O. K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors* 13–23 (2013).
59. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
60. Pedregosa, F. *et al.* *LinearSVC* (2024). Accessed: 2024-12-10.
61. Pedregosa, F. *et al.* *SGDClassifier* (2024). Accessed: 2024-12-10.
62. Pedregosa, F. *et al.* *Logistic Regression* (2024). Accessed: 24 December, 2024.
63. Weigand, H.-G. A discrete approach to the concept of derivative. *ZDM* **46**, 603–619 (2014).
64. Han, J., Kamber, M. & Pei, J. 3 - data preprocessing. In Han, J., Kamber, M. & Pei, J. (eds.) *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, 83–124, https://doi.org/10.1016/B978-0-12-381479-1.00003-4 (Morgan Kaufmann, Boston, 2012), third edition edn.
65. Pedregosa, F. *et al.* *StandardScaler — scikit-learn documentation* (2011). Accessed: 24 December, 2024.
66. Han, J., Kamber, M. & Pei, J. 8 - classification: Basic concepts. In Han, J., Kamber, M. & Pei, J. (eds.) *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, 327–391, https://doi.org/10.1016/B978-0-12-381479-1.00008-3 (Morgan Kaufmann, Boston, 2012), third edition edn.
67. Pedregosa, F. *et al.* *StratifiedKFold — scikit-learn documentation* (2011). Accessed: 24 December, 2024.
68. Adnan, Q. & A. B. M. Alim Al, I. RailVibes (2025). Available at: https://github.com/maqtanim/rail-vibes.

Author contributions statement

A.B.M.A.A.I. and A.Q. initiated the research idea. A.Q., W.J., M.R.S., S.I.R. and M.H.B. conducted the experiments and collected all the data. A.Q. and D.P. preprocessed the dataset. A.Q., and M.A.O.Z. analyzed the results. A.Q., M.A.O.Z., D.P., and J.N. prepared the manuscript. A.B.M.A.A.I. supervised the research as well as reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.Q.