

Article

Detecting Rail Surface Contaminants Using a Combined Short-Time Fourier Transform and Convolutional Neural Network Approach

Gerardo Hurtado-Hurtado, Tania Elizabeth Sandoval-Valencia , Luis Morales-Velázquez and Juan Carlos Jáuregui-Correa * 

Faculty of Engineering, Autonomous University of Queretaro, Santiago de Queretaro 76010, Mexico; gerardo.hurtado@uaq.mx (G.H.-H.); tania.sandoval@uaq.mx (T.E.S.-V.); luis.moralesv@uaq.mx (L.M.-V.)

* Correspondence: jc.jauregui@uaq.mx

Abstract

Condition monitoring of railway track surfaces is crucial for ensuring the safety, operational efficiency, and effective maintenance of railway systems. This work presents a data-driven modelling and an experimental methodology for identifying and classifying contaminants on railway tracks using vibration analysis and artificial intelligence techniques. In this study, the railway dynamics were physically simulated using a 1:20 scaled test rig, where the rails were treated with various contaminants (oil, water, and sand), and the resulting vehicle vibrations were recorded by on-board accelerometers and gyroscopes. To construct the predictive model, a hybrid architecture was designed integrating Short-Time Fourier Transform (STFT) for time-frequency feature extraction and a multi-channel Convolutional Neural Network (CNN) for pattern recognition. Initial results indicate that accelerometer data, particularly from longitudinal and lateral vibrations, are more effective than gyroscope data for classifying certain contaminants. To enhance classification robustness, this work introduces a multi-channel CNN that simultaneously processes the most informative signals, leading to a significant improvement in detection accuracy across all tested contaminants. This study validates the effectiveness of the proposed methodology as a robust and reliable solution for contaminant detection, while also confirming the utility of the scaled testbed as a valuable platform for future research in railway dynamics.

Keywords: railway condition monitoring; vibration analysis; artificial intelligence; convolutional neural networks; contaminant detection



Academic Editor: Wei Gao

Received: 23 November 2025

Revised: 8 January 2026

Accepted: 3 February 2026

Published: 9 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Monitoring rail contaminants and friction coefficients is of utmost importance in railway systems due to their significant impacts on safety, operational efficiency, maintenance costs, and overall vehicle performance [1,2]. Detecting rail contaminants is vital to ensuring proper acceleration and deceleration, especially during braking and starting.

The coefficient of friction directly affects the braking distance and stability of the train. Real-time information on rail friction conditions is optimal for enhanced control under any scenario or adverse condition. For example, high friction is required during starting and braking, while a low coefficient is desirable to prevent squealing, excessive wear, and ripples in both rails and wheels. A low friction coefficient also contributes to decreased energy consumption on curved sections, especially on very long high-speed trains [1,3].

Trains are exposed to the elements, making them susceptible to various natural and man-made contaminants, as well as numerous track irregularities. In these cases, the friction coefficient varies significantly as the train moves along its path, influenced by a multitude of environmental and operational parameters, and by contaminants present on the running surface, such as water, grease, and dry leaves [4]. The complexity of this interaction is evidenced by studies showing how environmental factors such as temperature and water content can significantly influence the mechanical properties of contact surfaces, even acting as lubricants and reducing drag [5,6]. The interactions, which originate at the wheel-rail contact, propagate to other train components through mechanical vibrations, including the ballast [7]. Contaminants and track irregularities are sources of such vibrations. A comprehensive understanding of these situations is crucial for sustaining optimal performance during acceleration and braking of railway vehicles, especially on inclined tracks.

Developing analytical or numerical models to predict wheel-rail interaction under contaminated conditions is mathematically complex. The presence of third-body materials (water, oil, sand) introduces highly non-linear tribological behaviors that are difficult to represent with traditional closed-form equations. Consequently, data-driven modelling emerges as a necessary alternative to capture these complex dynamic responses without relying on simplified physical assumptions.

Traditionally, the study of the effect of contaminants on railway dynamics has been carried out using classical experimental methods, such as pin-on-disc tribometers, twin-disc tribometers, portable tribometers and full-scale or reduced-scale test benches [8–11]. These methods allow simulating real conditions and studying the action of contaminants such as sand, leaves, water, oils and oxides on friction and wear [12–14]. Modern test benches integrate advanced sensors and data acquisition systems for real-time monitoring [15–17].

Future railway systems will rely on contaminant detection, integrated with advanced technologies, such as Artificial Intelligence (AI) and Machine Learning (ML). It will be the basis for smarter, more cautious, and sustainable operation in challenging environments. This concept aligns with the growing trend of digitalization in the railway industry, which aims to improve efficiency, reliability, and safety by implementing modern technologies and smart sensors [18]. To address the challenges and high costs associated with direct friction coefficient measurement in practical applications, future trains will implement these innovative solutions [19]. The implementation of AI and machine learning has enabled the development of predictive models and real-time friction coefficient estimation systems and contaminant detection [20,21]. AI-based methods enable real-time monitoring and prediction, but require large data volumes and experimental validation. The current trend is to integrate both approaches, using testbeds to validate AI models and onboard systems for continuous monitoring [16].

An artificial intelligence method has been developed, employing one of the most modern approaches for fault detection in mechanical systems. This AI algorithm first transforms vibration signals from train traction motors into time-frequency maps using the short-time Fourier transform (STFT), and then uses these maps as input to a convolutional neural network (CNN), which learns to identify both known and unknown faults, even in real-life, open-source environments. This method has demonstrated performance superior to that of traditional CNNs, significantly enhancing the accuracy of fault detection in various mechanical components, such as the classification of fault states in asynchronous motors [22], fault diagnosis of bearings in CNC machines [23], bearing fault detection using vibration signals [24], bearing fault detection under variable conditions [25], and rotor fault detection [26]. This method has also been implemented to detect faults in train engines [27]. CNNs are the state-of-the-art in computer vision and are very powerful

for image recognition, which can be applied directly alongside the STFT technique, as in this case, the spectrograms are processed as 2D images [28]. The combined application of STFT and CNN allows the simultaneous analysis of both temporal and frequency domain information of the signals, thereby facilitating the recognition of intricate patterns associated with different fault types.

The application of advanced sensor systems such as accelerometers and gyroscopes, in conjunction with AI algorithms and spectral analysis techniques such as STFT, is critical for the timely identification of faults and condition monitoring [29]. Nonetheless, no prior research has implemented this particular pair of techniques for the detection or classification of rail contaminants.

To address this gap, this paper details an experimental investigation into the identification of contaminants on railway rails through the application of vibration analysis and AI. The central hypothesis of this study is that the unique vibrational signatures generated by different surface contaminants (oil, water, and sand) can be effectively identified and classified by processing on-board sensor data through a combined Short-Time Fourier Transform (STFT) and Convolutional Neural Network (CNN) model. To test this hypothesis, a methodology was developed using a 1:20 scale test bench. The vibration data was collected with accelerometers and gyroscopes. Subsequently, the acquired data undergoes processing using STFT to generate spectrograms, which are then utilized as input for a 2D CNN. Performance is evaluated using Precision, Recall, and F1-Score metrics. A data augmentation technique is applied during the convolutional network training process to improve the robustness of this model. The objective is to create an artificial intelligence-based methodology for identifying and classifying contaminants in rails. The findings indicate that accelerometer data, particularly longitudinal and lateral vibrations, are significantly more effective than gyroscope data for this task. The model demonstrates exceptional accuracy in identifying sand and oil, but struggles to distinguish between wet and clean rails due to the similarity of their vibration signals. To overcome this limitation, an additional methodology integrating a multi-channel CNN is proposed. This adjustment significantly improves detection accuracy across all rail conditions, demonstrating that it is an effective solution for contaminant classification on railway tracks. The findings of this study also demonstrate the effectiveness and reliability of the 1:20 scale test rig, employed for data acquisition, in this particular domain of railway dynamics investigation. This study aims to validate this proposed methodology as an effective solution for contaminant classification on railway tracks.

2. Materials and Methods

2.1. Test Rig

The experimental rig is a single railcar built at a 1:20 scale, which sits on a pair of bogies. Tractive force is generated exclusively by the rear bogie, powered by a servomotor connected to the drive axle via a belt transmission, see Figure 1. This setup allows the vehicle to reach a maximum speed of 1.5 m/s on an unobstructed track. The wagon's body is connected to the bogies through a secondary suspension system composed of eight springs. Housed in the central part of the car are all the essential electronics, including the motor control unit, sensor circuitry, and the data acquisition system. The vehicle's main specifications are detailed in Table 1.

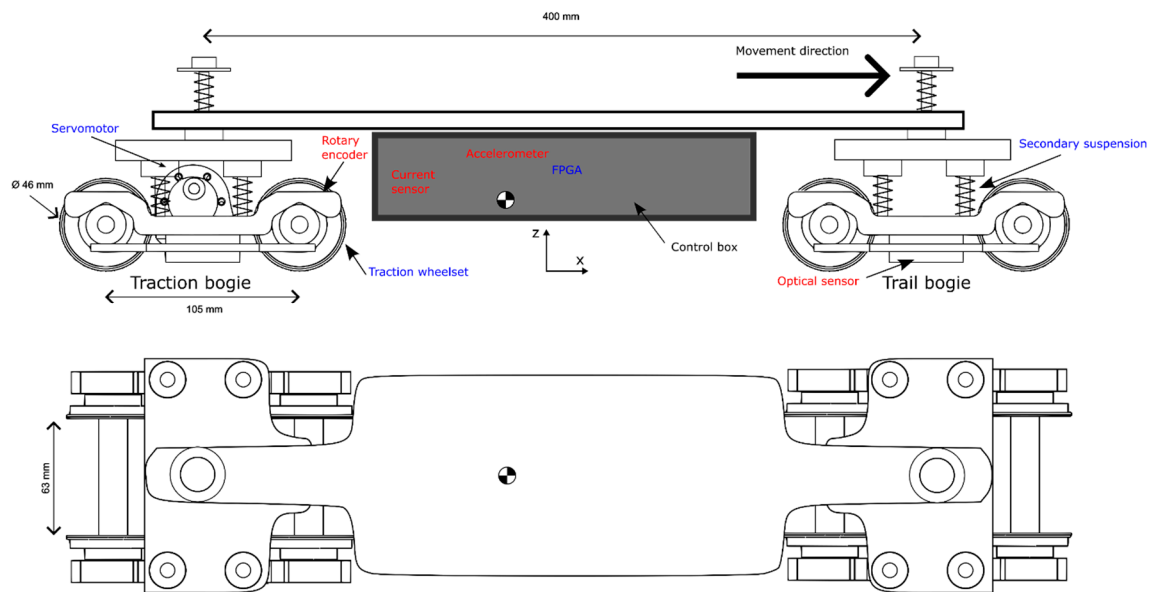


Figure 1. The dimensions and primary components of the model vehicle employed in the experiments, presented from both a lateral and an aerial perspective.

Table 1. Main vehicle parameters.

Physical Feature	Parameter	Value
Mass	Total mass	4.4 kg
	Traction bogie	1.0 kg
	Towed bogie	0.64 kg
	Vehicle body	2.76 kg
Dimensions	Total length	506 mm
	Total height	100 mm
	Total width	120 mm
	Wheel radius	23 mm
Suspension	Secondary	5.82 N/mm
Center of mass	Measured from reference frame	x: 42 mm
		y: 0 mm
		z: 62 mm

2.1.1. Testbed

The test bench where the railway system experiments are conducted is a horizontal closed circuit, similar to a β track, as shown in Figure 2. The testbed, featuring a steel frame with leveling feet and a plywood surface, is placed on a table. This table measures 2.36 m wide, 5.68 m long, and 0.92 m high. The track sleepers are machined into the plywood sheet surface. The rails consist of 5 mm diameter round-profile steel rods, which are mounted in longitudinal grooves on the sleepers and bonded with epoxy resin.



Figure 2. Panoramic view of the testbed. The rails are a β -shaped closed loop track.

2.1.2. Data Acquisition System

The vehicle monitoring and control system includes an FPGA board, sensors, and a servomotor, as shown in Figure 3. The FPGA board receives instructions and configurations from a Bluetooth module to control vehicle speed and stores the data on a microSD card. The control box contains the three-axis accelerometer and the three-axis gyroscope, mounted on a single chip, to monitor vehicle vibrations. The three accelerometers are aligned with the vehicle's three principal axes to measure longitudinal, lateral, and vertical vibrations, while the gyroscope measures the angular velocities, or rate of rotation, around these respective axes (roll, pitch, and yaw). The drive wheel speed is measured by a 600-pulse-per-revolution incremental rotary encoder mounted on the drive bogie, while an optical sensor mounted on the other (trailer) bogie measures the vehicle speed. Table 2 summarizes some characteristics of the acquisition system components.

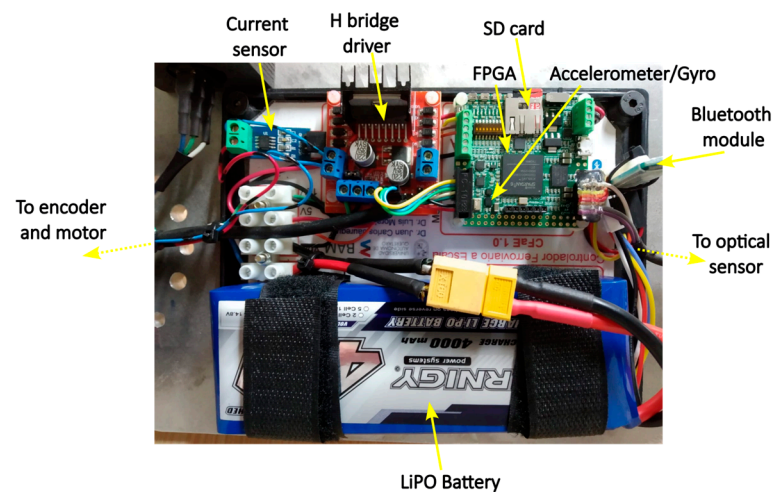


Figure 3. The acquisition system components inside the vehicle's control box.

Table 2. Key features of acquisition system components.

Component	Feature
Accelerometer	3-axis, MEMS sensor, model LSM6DS3, 4G
Gyroscope	3-axis, MEMS sensor, model LSM6DS3, 143°/s
Optical sensor	IR sensor, model TCRT5000
Rotary encoder	600 ppr, model DC5-24V 600.
Control card	DUA, Spartan 3.
Bluetooth	Bluetooth UART RS232, model HC-05.
Current sensor	Hall Effect, Model ACS712
LiPo battery	4000 mAh, 14.8 V
Servomotor	Pololu 37D with gearmotor, 12 V, 5.5 A, 12 W DC motor. The gear ratio is 19:1 with a maximum speed of 530 RPM and a torque of 0.83 N·m.
Servomotor driver	H-bridge, model L298N, 2 A, 30 V

2.2. Experiment

The test involves running the test railcar on a straight section of the track with contaminated rails and measuring vibrations using an accelerometer mounted in the center of the vehicle. The vibration data is stored on an SD card, mounted on the test vehicle's system, for subsequent analysis using a Python 3.11 program.

2.2.1. Contaminants and Their Application

The contaminants employed consisted of oil, water, and sand. Runs were also conducted on clean rails. For the oil contaminant, automotive transmission fluid was utilized. A uniform layer was applied along the entire straight section, ensuring that the rail sides remained uncontaminated. For water, a sprayer was used to completely soak the rail surface along the entire straight section. For the sand contaminant, fine grains (10 µm) were employed. The powder was dispersed by shaking the container, which facilitated the suspension of particles in the air and their subsequent precipitation onto the rails. Each of these contaminants reduces or increases the friction coefficient, as shown in Table 3. Each contaminant alters the coefficient of friction differently (as shown in Table 3), thereby generating characteristic vibration patterns, either stick-slip or damping, which the CNN model subsequently learns to distinguish.

Table 3. Rail contaminants and their order of application [30].

Order of Application	Contaminant	Coefficient of Friction
1	Clean rail	0.25
2	Water	0.1
3	Oil	0.01
4	Sand	0.5

2.2.2. Application of Contaminants

The test vehicle performed 10 runs for each contaminant. For oil, water, and sand, the rails were recontaminated every two runs, without altering the condition of the wheels. For the clean rail, the rails were not cleaned again until the tests were completed. Table 3 specifies the order in which the contaminants were applied.

Before applying each contaminant, the rails and wheels were cleaned with solvent (Thinner) and surgical gauze. They were thoroughly inspected to ensure the absence of residue from the previous contaminant and to avoid interference or noise in the data.

2.2.3. The Experiments

The vehicle begins its journey from the beginning of the straight section of the track (as shown in Figure 4); it accelerates and subsequently decelerates, coming to a complete stop precisely before the ensuing curve, whereupon it automatically resumes acceleration. It completes this same journey five times. Upon completing a contaminant, the vehicle's rails and wheels are cleaned, and the rails are subsequently recontaminated with the next substance.

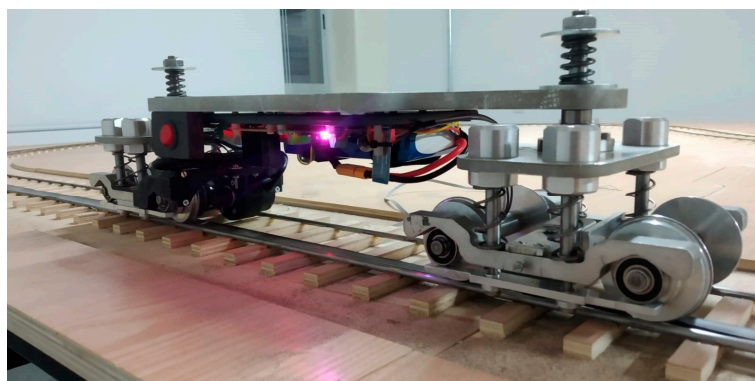


Figure 4. Photograph of the test vehicle located at the starting point of the test track.

The train ran at its maximum speed (1.5 m/s) in both directions, with either a polluting or a clean rail engaged individually on the straight section of the track. Measurements were taken of longitudinal (ax), lateral (ay), and vertical (az) accelerations, as well as roll (gx), pitch (gy), and yaw (gz) from the center of the test vehicle.

The vibrational data were collected utilizing an accelerometer (3 axes) and a gyroscope (3 axes). Samples were acquired over 0.5-s intervals, and the resulting data were stored for subsequent post-processing.

2.3. Data Processing

Multiple railcar runs were conducted for each of the four track conditions. From each run, once the train reached its maximum speed, two 0.5-s signal segments (500 samples at 1 kHz) were extracted. These segments were selected so that they did not overlap, with the first segment taken one second after reaching 95% of maximum speed and the subsequent segment immediately following the first; this guaranteed that the samples captured the characteristic dynamics of the condition and maintained statistical independence. Figure 5 provides a graphic flowchart illustrating this process.

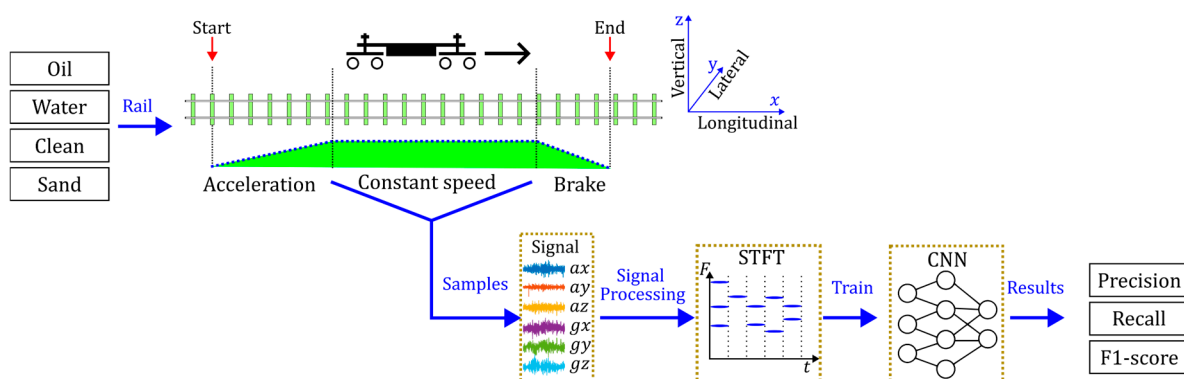


Figure 5. This diagram meticulously outlines the comprehensive experimental methodology employed in this study. It serves as a visual roadmap, illustrating each critical step and its interconnect-edness, ensuring clarity and reproducibility of the research process.

2.3.1. Processing Steps

1. Two 0.5-s samples were carefully collected from each train run. These samples were specifically selected when the train had already reached and was maintaining its maximum speed. This ensures that the captured data reflect stable operating conditions and are representative of train performance at its optimal state.
2. Once the samples were obtained, they were processed using the Short-Time Fourier Transform (STFT). This technique is essential for analyzing the frequency variation of a signal over time, which is crucial for identifying patterns and features in the sample data that might not be evident in the pure time domain. STFT enables spectrographic representation of data.
3. To improve the model's robustness and generalization capabilities, a data augmentation strategy was implemented. Data augmentation helps prevent overfitting and the training of a model that exhibits greater resilience to data variability.
4. A 2D CNN was used to analyze and classify the extracted features. CNNs are particularly effective at processing data with a grid structure, such as the spectrographic representations generated by the STFT. The 2D architecture allows the network to identify complex spatial and temporal patterns in the data, which is essential for accurate classification.

2.3.2. Short-Time Fourier Transform (STFT)

Spectrograms were generated from the 500 ms raw signals using STFT. A Tukey window was selected for this application due to its ability to mitigate discontinuities during signal segment analysis, as it maintains flatness in the center and tapers towards the edges. Additionally, the signals are overlapped by 50% to ensure that information at the edges of each window is not lost. This is common practice because it results in a smoother, more continuous spectrogram. The window size was defined as 64 samples. Since the sampling rate is 1000 Hz, this corresponds to a 64 ms time window. The 1 kHz sampling is adequate, in this instance, for capturing the natural frequencies of the vehicle and the wheel-rail contact dynamics within this scale model. The 64-sample window size offers a good balance between both frequency and time resolution. The STFT is defined as follows:

$$STFT = \{x[n]\} (m, \omega) \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n] w(n-m) e^{-j\omega n}, \quad (1)$$

where $x[n]$ is the accelerometer signal, $w[n]$ is the window function, m is the time, ω is the frequency.

The Tukey window is formally defined piecewise across its N discrete points. Its shape depends on a parameter α (alpha), which controls how wide the smoothed edges are. The formula for a point n (where n ranges from 0 to $N-1$) is:

$$w(n) = f(x) = \begin{cases} \frac{1}{2} \left[1 - \cos \left(\frac{2\pi n}{\alpha(N-1)} \right) \right], & 0 \leq n \leq \frac{\alpha(N-1)}{2} \\ 1, & \frac{\alpha(N-1)}{2} < n < (N-1) \left(1 - \frac{\alpha}{2} \right) \\ \frac{1}{2} \left[1 - \cos \left(\frac{2\pi(N-1-n)}{\alpha(N-1)} \right) \right], & (N-1) \left(1 - \frac{\alpha}{2} \right) \leq n \leq (N-1) \end{cases}, \quad (2)$$

where $w(n)$ is the window value at point n , N is the total number of points in the window (the STFT window size), n is the current index, ranging from 0 to $N-1$, and α is the shape parameter or taper ratio. This parameter quantifies the proportion of the window encompassed by cosine edges.

The Tukey window helps mitigate spectral leakage, a common problem when using a rectangular window, as it can cause a strong frequency to contaminate nearby frequencies. By smoothing the edges using the parameter α , the Tukey window reduces this leakage,

but by maintaining a flat center portion, it does not lose as much frequency resolution as a full cosine window would. It is ideal for analyzing signals that may have both transient and stationary components.

2.3.3. Data Augmentation

To avoid overfitting due to the limited number of samples (160 files), a data augmentation technique was implemented directly on the spectrograms.

1. Frequency Masking: A frequency band is randomly hidden. In this case, a 10 Hz band was used.
2. Time Masking: A time segment is randomly hidden. In this case, a width of 15 was used, corresponding to 15 ms.

With this method, the model learns more general and robust characteristics of each contaminant's vibrational fingerprint, thus improving its ability to detect unseen situations and making it more resilient to noise.

By forcing the model to classify a signal even when a portion of its frequency or time data is missing, it is taught not to rely on a single range to make its decision. This makes it more resilient to variations or anomalies in the real data.

2.3.4. 2D Convolutional Neural Network (CNN) Architecture

The CNN architecture was designed to extract hierarchical features from the 2D spectrograms generated by the STFT. The layer sequence is as follows, as depicted in Figure 6:

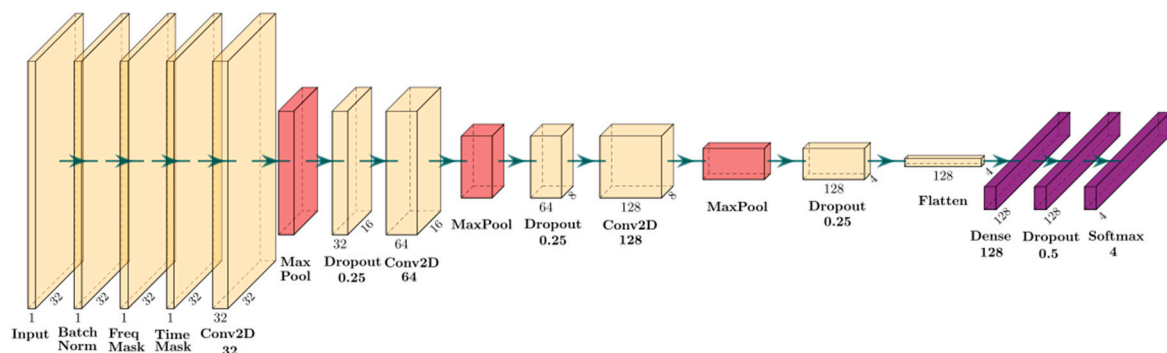


Figure 6. The architecture of the 2D Convolutional Neural Network (CNN) model is a meticulously designed system comprising multiple internal layers, each serving a distinct purpose in feature extraction and classification.

The input layer is designed to process 2D image data (STFT spectrograms). The input shape for a single channel (to consider one signal at a time) is (height, width, 1).

1. Next is Batch Normalization, which normalizes the activations from the previous layer to speed up training and improve stability.
2. Data Augmentation Layers. The Freq Mask and Time Mask layers, described above, are only activated during training.
3. The Convolutional Block 1 starts with a Conv2D layer (32 filters, 3×3 kernel, ReLU activation), followed by MaxPooling2D (2×2 window) to reduce dimensionality and complexity. Finally, Dropout (25%) prevents overfitting, promoting robust features.
4. The Convolutional Block 2 uses a Conv2D layer (64 filters, 3×3 kernel, ReLU) to extract features. Then, MaxPooling2D (2×2 window) reduces dimensionality and improves robustness. Finally, Dropout (25%) prevents overfitting, forcing the network to learn diverse representations.

5. The third convolutional block of the neural network includes a Conv2D layer with 128 3×3 filters and ReLU activation, followed by a 2×2 MaxPooling2D layer to reduce dimensionality and a 25% Dropout layer to prevent overfitting.
6. Finally, the classifier uses a Flatten layer to convert the two-dimensional feature maps into a one-dimensional vector. This is followed by a Dense layer of 128 neurons with ReLU activation. To prevent overfitting, it incorporates a 50% Dropout layer. The output layer is another Dense layer with four neurons and Softmax activation, which generates a probability distribution for the classes.

This organization of three main convolutional blocks (32-64-128) is a specific design to extract progressively more complex features. This configuration follows a standard design pattern known as VGG-style, which is widely used for feature extraction in computer vision tasks [31]. The progression of the filters in these blocks is key:

- The first Block starts with 32 filters, optimized to capture low-level features.
- The second Block increases to 64 filters, allowing the detection of medium-complexity patterns.
- The third Block culminates with 128 filters, focused on the identification of high-level patterns and more abstract semantic representations.

This VGG-style approach was meticulously chosen for this specific dataset, seeking an optimal balance between network depth (necessary for representation capacity) and computational cost, which ensures reasonable training and execution times.

The fundamental component of this architecture is the kernel or filter, a small matrix of weights that is passed through the input to generate feature maps. Each kernel is designed to detect specific patterns. Its main advantage is that its values are not predefined, but are learned by the model during training through the backpropagation algorithm. This process allows the network to specialize in identifying the most discriminatory time-frequency features for each type of pollutant. The use of kernels also allows for parameter sharing, making the model computationally efficient and capable of learning a hierarchical representation of the data, from simple edges in the first layers to complex textures in the deeper layers.

2.3.5. Model of CNN for a Single Input Channel (Monochannel)

The input is the 2D spectrogram resulting from the STFT. The convolution operation to generate an O^k output feature map is defined as:

$$O_{i,j}^k = \sigma \left(\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{i+m,j+n} \cdot K_{m,n}^k + b^k \right), \quad (3)$$

where I is the input spectrogram, K^k is the k -th convolution kernel, b^k is the bias term for the k -th kernel, O^k is the k -th output feature map, σ is the non-linear activation function, i, j are the spatial indices that traverse the height and width of the spectrogram, c is the index that traverses the input channels and M, N are the dimensions (height and width) of the kernel.

To compute the value at position (i, j) of the k th output map ($O_{i,j}^k$), the kernel slides over a region of the input I . The double summation represents the dot product operation between the kernel values and the corresponding values of the input spectrogram over which it is positioned. To this result is added a single bias term b^k , which allows the filter to learn an offset. Finally, the entire result is passed through the activation function σ , which introduces the non-linearity necessary for the network to learn complex patterns.

2.3.6. CNN Multichannel

The CNN Multichannel formula expands to include a sum over the channel dimensions:

$$O_{i,j}^k = \sigma \left(\sum_{c=0}^{C_{in}-1} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{i+m,j+n,c} \cdot K_{m,n,c}^k + b^k \right), \quad (4)$$

In addition to the double summation in (3), in this case, there is also a third summation that iterates over each of the input channels. For each channel c , a separate 2D convolution is performed between channel c of the input $I_{...,c}$ and channel c of the kernel $K_{...,c}^k$. Finally, the convolution results for the c channels are summed together. The single bias term b^k is added to this total sum. This combined result passes through the activation function σ .

2.3.7. CNN Training Process

The training of the model was systematically conducted using a precise set of hyperparameters and settings, which are detailed in Table 4.

Table 4. The hyperparameters and configurations utilized during the training process.

Parameter	Configuration/Value	Description
Optimizer	Adam	This optimization algorithm is efficient and adaptively adjusts the learning rate for each parameter.
Learning Rate	0.001	Robust and commonly used initial value for the Adam optimizer.
Number of Epochs	50	The model was trained for 50 full passes through the entire training dataset.
Batch Size	16	The model updates its weights after processing 16 spectrograms.
Loss Function	cross-entropy	It is best suited for multiclass classification problems with integer labels

2.3.8. Data Division and Preparation

To ensure an unbiased evaluation of model performance, the total dataset (160 files) was split as follows:

1. The data were divided into a training set (75%, 120 files) and a test set (25%, 40 files).
2. Before splitting, the entire data set was randomly shuffled to avoid any bias introduced by the order of data collection. A fixed seed was used to ensure the split was always the same, allowing for reproducible results.
3. The split was performed in a stratified manner. This ensures that the proportion of samples from each class (Water, Oil, Sand, Clean) is identical in both the training and test sets. This is a crucial step for small or unbalanced datasets.
4. During the model training process, performance was monitored at each epoch. To achieve this, the test data set was used as the validation set. This practice allows us to visualize how the model generalizes to the data in real time, which in turn facilitates early detection of overfitting.

2.3.9. Model Evaluation Metrics

To enrich this research, classifications are compared across directions and rotations; that is, the results show the accuracy of the ax , ay , and az accelerometers compared to the accuracy of the gx , gy , and gz gyroscopes. The evaluation results of the CNN model are presented using precision, recall, and F1-score metrics. These metrics are calculated for each situation, offering a comprehensive perspective on the CNN model's performance, which a solitary metric such as precision cannot adequately provide, especially when the data is unbalanced. In other words, it is used to understand how many times the

model is correct or wrong. Precision measures the quality of positive predictions using the following formula:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}, \quad (5)$$

High precision means the model has a low false positive rate. When the model says something is positive, it is most likely true. Recall, on the other hand, measures the model's ability to find all positive samples. In other words, it refers to the proportion of true positives the model was able to identify. Recall is calculated using the following formula:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}, \quad (6)$$

A high recall value indicates that the model has a low false negative rate, indicating the model's efficacy in accurately identifying positive instances. The F1-score value is the harmonic mean of the recall and precision:

$$\text{F1-score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (7)$$

In this case, it is better to use the harmonic mean instead of a simple average because it severely penalizes extreme values. A high F1-score means both precision and recall are high, indicating few false positives or false negatives.

3. Results

The Tukey window was used with an $\alpha = 0.25$, a length of 64 samples (equivalent to 64 ms), and a 50% overlap (32 samples) between adjacent segments. This configuration achieves a spectral resolution of 15.625 Hz. This resolution represents a balance between precision in the time and frequency domains, optimizing the ability to identify vibratory patterns. The $\alpha = 0.25$ value is commonly used in mechanical measurements. This is a standard choice because it avoids the spurious frequencies that arise from sharp transitions, while still maintaining a temporal resolution that avoids the final outcome being overly compromised by excessive smoothness.

Figure 7 displays spectrograms of the accelerometer data, and Figure 8 presents those for the gyroscope data, comparing the most extreme conditions: rails with oil (low friction) and rails with sand (high friction). This representation illustrates the energy distribution of the signals in the two-dimensional plane of time and frequency. Specifically, by moving from the purely temporal domain to this combined time-frequency domain, it is possible to accurately identify the presence and dynamic behavior of the distinct harmonic or spectral components at each measurement instant. This capability is crucial for the diagnosis and understanding of non-stationary phenomena, where the characteristics of the signal vary over time.

The model's performance, which was trained with vibration data (STFT results from accelerations and rotations) in all directions to detect the different test contaminants in the rails, is shown through the aforementioned metrics. The results are presented in bar graphs, grouped according to the different contaminants (Oil, Water, Clean, and Sand). Presenting these three metrics as indicators of the model's performance clearly demonstrates the true role of each sensor. In other words, this analysis enables precise identification of sensors contributing to contaminant detection, distinguishing them from those introducing noise or misclassification.

The precision, recall, and F1-score metrics are shown in Figures 9–11, respectively. In these graphs, it is easy to see that the accelerometer data (blue bars) is much more effective than the gyroscope data (red bars). Specifically, the ax and ay accelerometers (sensitive in

the longitudinal and lateral directions, respectively) consistently emerge as the most robust predictors across virtually all scenarios.

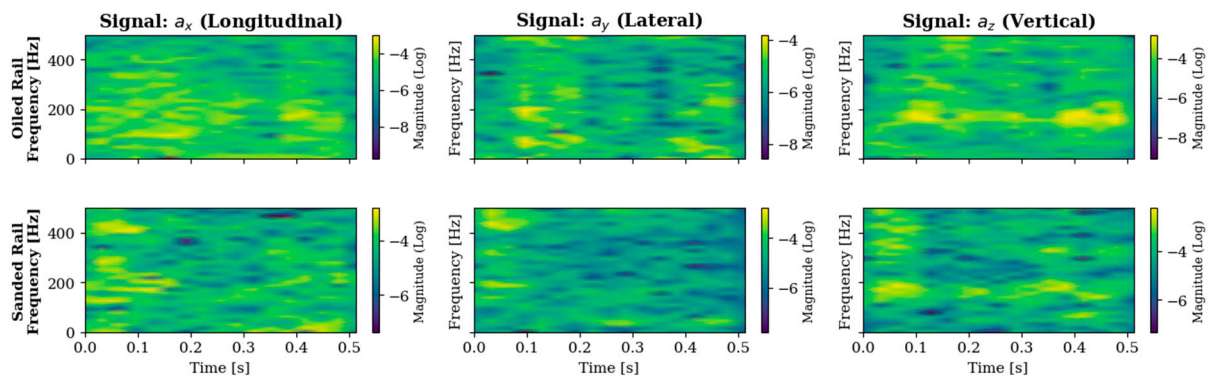


Figure 7. STFT-spectrograms obtained from the a_x , a_y , and a_z accelerometer signals, showing the differences between oiled rails versus sanded rails.

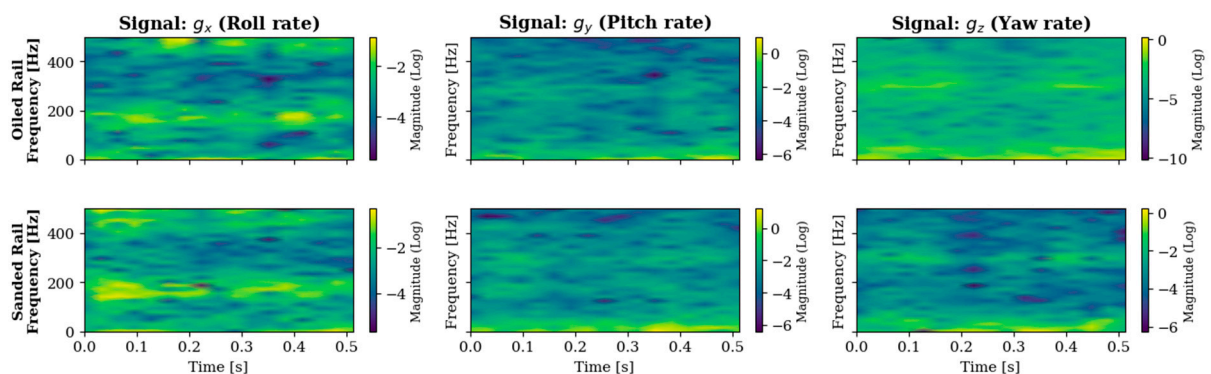


Figure 8. STFT-spectrograms obtained from the g_x , g_y , and g_z gyro signals, showing the differences between oiled rails versus sanded rails.

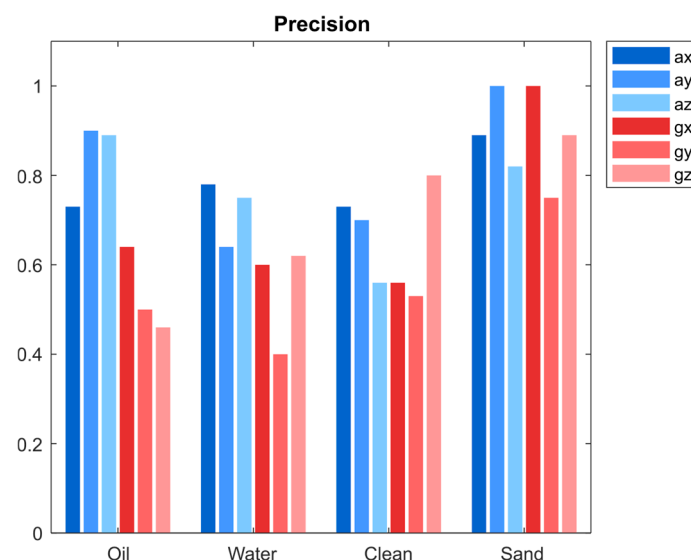


Figure 9. The single-channel CNN model's precision for each contaminant on the rails is graphically depicted, showcasing all directions and rotations. Sand demonstrates the highest precision, indicating a low false positive rate for this particular contaminant.

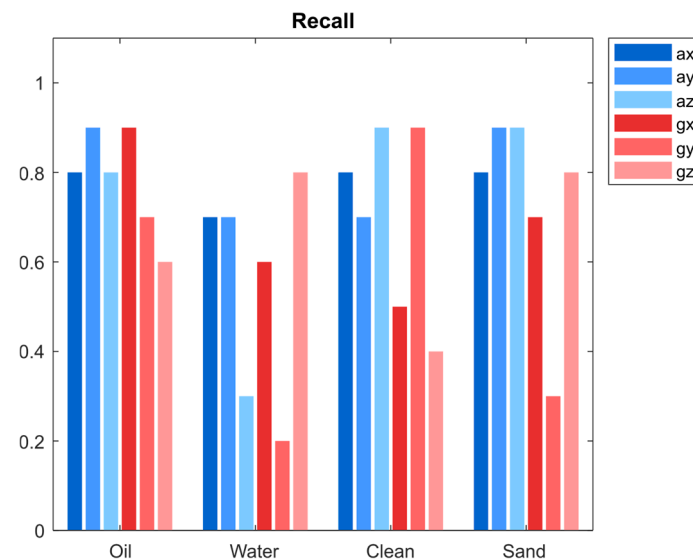


Figure 10. The recall of the mono-channel CNN model for each rail contaminant, across all directions and rotations. Oil and sand consistently demonstrate high recall, indicating a low false-negative rate, particularly in the accelerometer data.

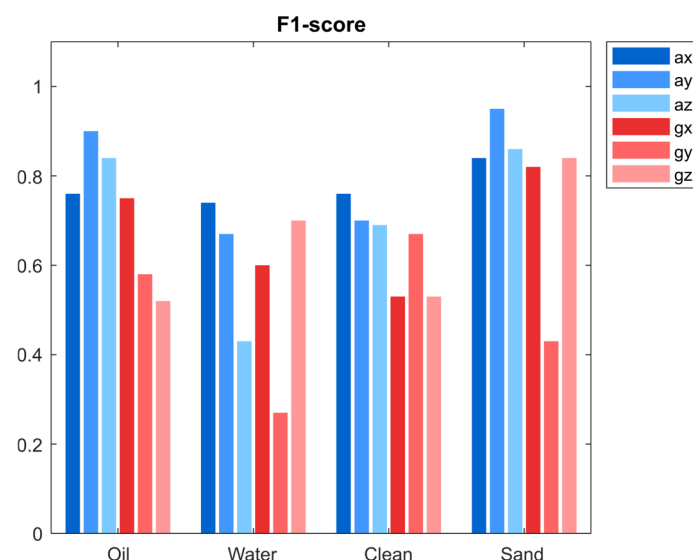


Figure 11. Overall F1-Score for all contaminants and sensors of the mono-channel CNN model. The accelerometer data has the best performance, meaning it indicates few false positives and false negatives.

Gyroscope data, particularly from the gy and gz axes, exhibit poor performance in distinguishing between wet and clean rails. The low values for Recall and F1-score suggest that the rotations in these directions do not substantially contribute to the model's predictive accuracy and may even introduce inaccuracies, in relation to these contaminants. The only partial exception is the gx gyroscope, which seems to have some relevance, although still inferior to the accelerometer scores.

Improving the CNN Mono-Channel Model

After evaluating the classification performance of each sensor signal individually, this section presents an analysis using the three accelerometer signals, given their superior performance in the preceding analysis. Then, to exploit the synergistic information between the different acceleration motions, a CNN model capable of simultaneously processing the three signals (ax , ay , and az) was implemented. This multi-channel approach was

designed to capture the complex correlations and patterns emerging from the interaction of vibrations along the three spatial directions, leading to a significant improvement in the accuracy and robustness of the track condition contaminant classification.

The CNN model remains consistent with the previous architecture. The only difference is how the data is prepared and presented before entering this neural network. Unlike the single-channel approach, where each input sample was a 2D spectrogram, in the multi-channel model, a spectrogram is generated for each accelerometer signal. Subsequently, these three spectrograms are stacked along a new channel dimension, yielding a singular three-dimensional input tensor.

The input to the CNN had a shape of [height, width, 1], where the trailing '1' indicates a single channel of information. This is analogous to a grayscale image. Now, in this multi-channel approach, the input shape changed to [height, width, 3], where the trailing '3' represents the three spectrograms of x , ay , and az , resulting from the STFT, stacked together. This is analogous to an RGB color image, where each color channel (red, green, blue) represents a spectrogram corresponding to one axis of the accelerometer.

While the CNN model used is identical, its internal behavior changes due to the new input shape. While in the single-channel model, the Conv2D layer kernels were essentially 2D and operated on the single input channel to extract features, in the multi-channel model, the Conv2D layer kernels are now 3D (they have a depth of 3, to match the three input channels). With these changes, in addition to performing convolutions that detect spatial (time-frequency) patterns in each spectrogram, the network also learns correlation patterns between the three axes simultaneously, significantly improving predictions.

Below are the outcomes obtained with the multi-channel model, as shown in Figures 12–14. It should be noted that the exact same dataset was used to train and evaluate this model. The Precision, Recall, and F1-score metrics from this new analysis reveal a quantitative improvement, further indicating that the signal pattern for each contaminant is more distinctly characterized by the entire motion vector, encompassing all three directions, and to a lesser extent by its individual components.

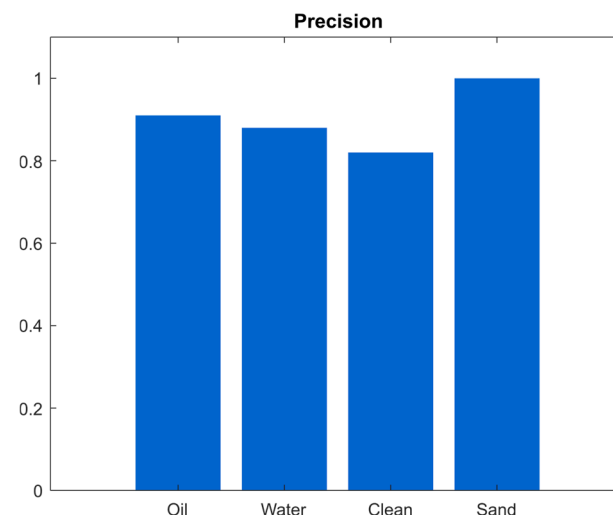


Figure 12. The precision of the multi-channel CNN model for each contaminant on the rails is graphically depicted. Almost all contaminants exhibit a low false positive rate, with sand being the most readily classifiable contaminant.

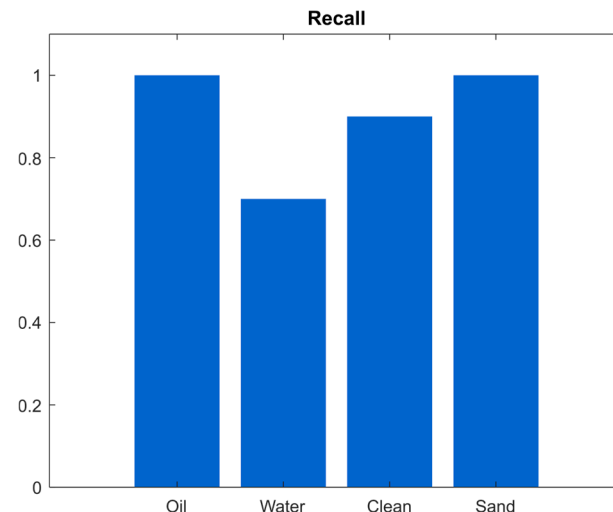


Figure 13. The recall of the multi-channel CNN model for each rail contaminant, across all directions and rotations. Nearly all contaminants enhanced the recall metric with this model, indicating a low false-negative rate. Oil and sand continue to exhibit the highest recall.

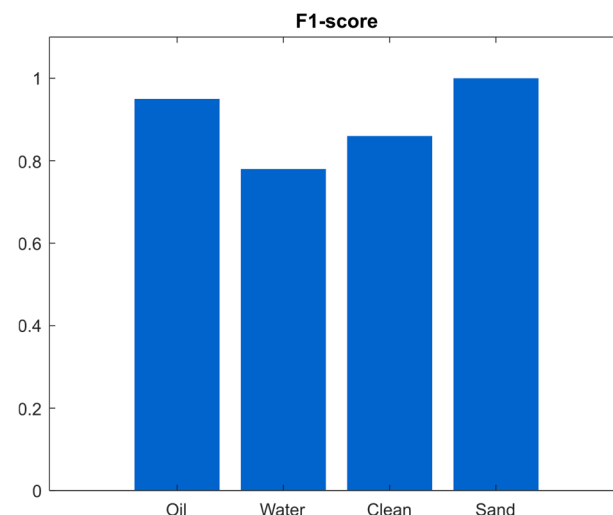


Figure 14. The multi-channel CNN model demonstrates exceptionally high performance across all contaminants and sensors, as evidenced by the overall F1-Score. This chart underscores the effectiveness of the multi-channel approach in achieving strong results for every contaminant.

4. Discussion

4.1. Single-Channel CNN Model Discussion

The graphs clearly show that accelerometers ax , and ay are the most valuable sensors. They have high accuracy and high recall for almost all classes. This means that longitudinal and lateral vibrations are the key physical characteristics for differentiating between contaminants such as oil, sand, water, and clean rails.

Regarding the contaminants, the STFT-CNN-based model demonstrates high performance in detecting sand and oil on the rails. The F1-score, Precision, and Recall are high, especially with the accelerometer data. This indicates that both sand and oil generate very distinctive signals that are easy for the model to classify.

Also, when the rail is clean, the model can identify it with acceptable efficiency using the accelerometer, but it is clearly more difficult than with sand or oil. The difficulty in distinguishing between wet and clean rails is likely attributable to the wheel-rail contact potentially persisting in a boundary lubrication regime where the water film is displaced by

the vehicle's weight, resulting in significant asperity contact, similar to the dry condition. This would explain the lack of a distinct vibration signature compared to the high-friction sand or viscous oil scenarios.

Water is the most problematic contaminant. While the accelerometers ax and az achieve decent performance, the gyroscope is practically useless here. The low overall score suggests that water contamination produces the least distinctive signal or one that closely resembles another state (probably very similar to the signal from a clean rail).

4.2. Multi-Channel CNN Model Discussion

The improvement reached with this model is evident, particularly under previously ambiguous conditions, such as wet rail and clean rail. For example, the previous single-channel model yielded a low Recall and F1-score (approximately 0.4–0.7) for wet rail, whereas the multi-channel model achieved an F1-score approaching 0.78. Although this class continues to exhibit the lowest metric, the improvement is substantial. This model no longer relies on a singular axis. Instead, it utilizes joint information to mitigate ambiguity. The model also improved by more easily identifying clean rail, consistently reaching close to 0.85 in the F1-score. There was even a significant improvement in identifying the oil over the rails, reaching an F1-score of 0.95, demonstrating classification with few errors.

The F1-score for sand identification was already high when using longitudinal vibrations (ax accelerometer), but when using all accelerometers, the Recall and F1-scores reach almost 1.0 (or 100%). The model is able to identify this contamination without error, thanks to its vibration signature in all three directions.

5. Conclusions

This paper developed and validated an artificial intelligence-based method for the automatic detection of contaminants at the wheel-rail interface. Using vibration data from a scaled model, we demonstrated that a multi-channel Convolutional Neural Network (CNN), fed with spectrograms from accelerometer signals, can classify track conditions with high accuracy. The results confirm that this approach not only identifies the presence of contaminants such as oil, water, and sand but also effectively differentiates between them, overcoming the limitations of single-channel data analysis.

This study serves as a successful proof-of-concept, revealing the remarkable potential of vibration analysis and deep learning as tools for real-time monitoring of railway infrastructure. The ability to reliably and automatically detect adverse track conditions marks a significant step toward enhancing operational safety, optimizing predictive maintenance, and enabling the autonomous railway systems of the future. The findings from this research provide a solid methodological foundation for the crucial next step: validating and adapting this technology for implementation on full-scale trains.

Limitations of the Study and Future Work Prospects

While the multi-channel STFT-CNN method demonstrated efficacy in laboratory settings for identifying rail contaminants, certain limitations persist. The 1:20 scale model employed may not fully replicate the dynamics of actual train operations, and the experimental environment does not adequately account for real-world variables such as ambient noise, meteorological conditions, track degradation, or fluctuations in train speed and weight. The experimental design was meticulously structured to constrain the scope of the analysis to a defined set of controlled conditions, thereby facilitating the precise isolation and rigorous evaluation of the contaminating variable's effect. Furthermore, the scope of contaminants investigated was limited to oil, water, and sand, thereby excluding other significant low-adhesion agents like ice, snow, grease, or wet leaves. Additionally, the

current study is strictly limited to classifying the contaminant type, without providing a quantitative estimation of the severity or amount present on the rails. These areas warrant further investigation in subsequent research.

The promising results of this study open new avenues for future research to advance the technology towards practical implementation. A critical next step is to validate the methodology on a full-scale, in-service train using industrial-grade accelerometers. Future work should also focus on creating a more comprehensive database by testing a wider range of contaminants and operating conditions. Furthermore, performance could be enhanced by exploring more advanced deep learning architectures and explicitly benchmarking the proposed method against traditional machine learning baselines, such as Support Vector Machines (SVM) and Random Forest (RF). Finally, achieving real-time, on-board deployment will require investigating model optimization and suitable hardware.

Author Contributions: G.H.-H. performed the analysis, interpreted the data, and wrote the paper. G.H.-H. and T.E.S.-V. conceived and designed the experiment. L.M.-V. made the instrumentation and the railway vehicle for data acquisition and contributed to the design of the AI algorithm. J.C.J.-C. revised and gave final approval of the version to be submitted and all subsequent versions. All authors have read and agreed to the published version of the manuscript.

Funding: The authors declare that they received financial support from the Universidad Autónoma de Querétaro during the preparation of this manuscript.

Data Availability Statement: The original contributions presented in the study are included on the Kaggle page with the name “RailwayUAQ”, or more queries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank CONAHcyT for supporting this work in 2023 via the project on Frontier Science with the number CF-2023-I-204, and for supporting the postdoctoral stays in Mexico 2023 (1).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rahaman, M.L.; Bernal, E.; Spiriyagin, M.; Bosomworth, C.; Sneath, B.; Wu, Q.; Cole, C.; McSweeney, T. An investigation into the effect of slip rate on the traction coefficient behaviour with a laboratory replication of a locomotive wheel rolling/sliding along a railway track. *Tribol. Int.* **2023**, *187*, 108773. [\[CrossRef\]](#)
2. Zhao, Y.; Liang, B.; Iwnicki, S. Friction coefficient estimation using an unscented Kalman filter. *Veh. Syst. Dyn.* **2014**, *52*, 220–234. [\[CrossRef\]](#)
3. Zirek, A.; Onat, A. A novel anti-slip control approach for railway vehicles with traction based on adhesion estimation with swarm intelligence. *Railw. Eng. Sci.* **2020**, *28*, 346–364. [\[CrossRef\]](#)
4. Wu, B.; Xiao, G.; An, B.; Wu, T.; Shen, Q. Numerical study of wheel/rail dynamic interactions for high-speed rail vehicles under low adhesion conditions during traction. *Eng. Fail. Anal.* **2022**, *137*, 106266. [\[CrossRef\]](#)
5. Wang, Y.; Zhao, Y.; Mao, X.; Yin, S. Impact of Climate Change on the Performance of Permafrost Highway Subgrade Reinforced by Concrete Piles. *Future Transp.* **2023**, *3*, 996–1006. [\[CrossRef\]](#)
6. Al-Maliki, H.; Meierhofer, A.; Trummer, G.; Lewis, R.; Six, K. A new approach for modelling mild and severe wear in wheel-rail contacts. *Wear* **2021**, *476*, 203761. [\[CrossRef\]](#)
7. Moreno, A.G.; López, A.A.; Carrasco García, M.G.; Turias, I.J.; Ruiz Aguilar, J.J. A novel application of computational contact tools on nonlinear finite element analysis to predict ground-borne vibrations generated by trains in ballasted tracks. *Modelling* **2024**, *5*, 1454–1468. [\[CrossRef\]](#)
8. Singh, R.; Shindhe, M.; Rawat, P.; Srivastava, A.; Singh, G.; Verma, R.; Bhutto, J.; Hussein, H. The Effect of Various Contaminants on the Surface Tribological Properties of Rail and Wheel Materials: An Experimental Approach. *Coatings* **2023**, *13*, 560. [\[CrossRef\]](#)
9. Khalladi, A.; Elleuch, K. Tribological Behavior of Wheel–Rail Contact Under Different Contaminants Using Pin-On-Disk Methodology. *J. Tribol.-Trans. ASME* **2017**, *139*, 011102. [\[CrossRef\]](#)
10. Zani, N.; Petrogalli, C.; Battini, D. Optimizing Railway Tribology: A Systematic Review and Predictive Modeling of Twin-Disc Testing Parameters. *Lubricants* **2024**, *12*, 382. [\[CrossRef\]](#)
11. Zhu, Y.; Yang, H.; Wang, W. Twin-disc tests of iron oxides in dry and wet wheel–rail contacts. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2016**, *230*, 1066–1076. [\[CrossRef\]](#)

12. Olofsson, U.; Lyu, Y.; Zhu, Y. Mapping the friction between railway wheels and rails focusing on environmental conditions. *Wear* **2015**, *324*, 122–128. [[CrossRef](#)]
13. Zhu, Y.; Olofsson, U.; Nilsson, R. A field test study of leaf contamination on railhead surfaces. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2014**, *228*, 71–84. [[CrossRef](#)]
14. Gallardo-Hernández, E.; Lewis, R. Twin disc assessment of wheel/rail adhesion. *Wear* **2008**, *265*, 1309–1316. [[CrossRef](#)]
15. Valena, M.; Omasta, M.; Kvarda, D.; Galas, R.; Krupka, I.; Hartl, M. An approach for the creep-curve assessment using a new rail tribometer. *Tribol. Int.* **2023**, *191*, 109153. [[CrossRef](#)]
16. Harmon, M.; Santa, J.; Jaramillo, J.; Toro, A.; Beagles, A.; Lewis, R. Evaluation of the coefficient of friction of rail in the field and laboratory using several devices. *Tribol.-Mater. Surf. Interfaces* **2020**, *14*, 119–129. [[CrossRef](#)]
17. Krishnan, J.; Yang, Z.; Li, Z. Wheel/Rail Adhesion and Coefficient of Friction Measurement using Downscaled Test Rig. In *Proceedings of the Sixth International Conference on Railway Technology, Prague, Czech Republic, 1–5 September 2024*; Civil-Comp Press: Edinburgh, UK, 2024. [[CrossRef](#)]
18. Morin, X.; Olsson, N.O.; Lau, A. Managerial challenges in implementing European rail traffic management system, remote train control, and automatic train operation: A literature review. *Future Transp.* **2024**, *4*, 1350–1369. [[CrossRef](#)]
19. Yin, N.; Yang, P.; Liu, S.; Pan, S.; Zhang, Z. AI for tribology: Present and future. *Friction* **2024**, *12*, 1060–1097. [[CrossRef](#)]
20. Zhao, Y.; Shen, L.; Jiang, Z.; Zhang, B.; Liu, G.; Shu, Y.; Peng, B. Real-time wheel–rail friction coefficient estimation and its application. *Veh. Syst. Dyn.* **2023**, *61*, 2598–2612. [[CrossRef](#)]
21. Onat, A.; Voltr, P.; Lata, M. A new friction condition identification approach for wheel–rail interface. *Int. J. Rail Transp.* **2017**, *5*, 127–144. [[CrossRef](#)]
22. Wang, L.; Zhao, X.; Wu, J.; Xie, Y.; Zhang, Y. Motor Fault Diagnosis Based on Short-time Fourier Transform and Convolutional Neural Network. *Chin. J. Mech. Eng.* **2017**, *30*, 1357–1368. [[CrossRef](#)]
23. Iqbal, M.; Madan, A. CNC Machine-Bearing Fault Detection Based on Convolutional Neural Network Using Vibration and Acoustic Signal. *J. Vib. Eng. Technol.* **2022**, *10*, 1613–1621. [[CrossRef](#)]
24. Zhang, Q.; Deng, L. An Intelligent Fault Diagnosis Method of Rolling Bearings Based on Short-Time Fourier Transform and Convolutional Neural Network. *J. Fail. Anal. Prev.* **2023**, *23*, 795–811. [[CrossRef](#)]
25. Pham, M.; Kim, J.; Kim, C. Accurate Bearing Fault Diagnosis under Variable Shaft Speed using Convolutional Neural Networks and Vibration Spectrogram. *Appl. Sci.* **2020**, *10*, 6385. [[CrossRef](#)]
26. Jung, H.; Choi, S.; Lee, B. Rotor Fault Diagnosis Method Using CNN-Based Transfer Learning with 2D Sound Spectrogram Analysis. *Electronics* **2023**, *12*, 480. [[CrossRef](#)]
27. Tong, R.; Xu, Z.; Qu, H.; Jiang, K. Kernel Based Open CNN Algorithm for Known and Unknown Fault Diagnosis of Train Traction Motor. In *Proceedings of the 2024 Global Reliability and Prognostics and Health Management Conference (PHM-Beijing), Beijing, China, 1–13 October 2024*; IEEE: New York, NY, USA, 2024; pp. 1–7.
28. Bu, Q.; Lyu, P.; Sun, R.; Jing, J.; Lyu, Z.; Hou, S. Fault Diagnosis Method Using CNN-Attention-LSTM for AC/DC Microgrid. *Modelling* **2025**, *6*, 107. [[CrossRef](#)]
29. Gonzalez-Jorge, H.; Ríos-Otero, E.; Aldao, E.; Balvís, E.; Veiga-López, F.; Fontenla-Carrera, G. Rail Maintenance, Sensor Systems and Digitalization: A Comprehensive Review. *Future Transp.* **2025**, *5*, 83. [[CrossRef](#)]
30. Hurtado-Hurtado, G.; Morales-Velazquez, L.; Otremba, F.; Jáuregui-Correa, J.C. Railcar Dynamic Response during Braking Maneuvers Based on Frequency Analysis. *Appl. Sci.* **2023**, *13*, 4132. [[CrossRef](#)]
31. Khan, A.; Sohail, A.; Zahoor, U.; Qureshi, A. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2019**, *53*, 5455–5516. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.