Enhancing Wagon Fault Detection Using Acoustic Signals and Deep Transfer Learning: From Scratch to Full Fine-Tuning

Jordana Reis[1], Rhuan Garcia[2], and Flávio Varejão[3]

[1] *Departamento de Tecnologia, Vale S.A., Vitória, ES, 29.090-860, Brazil*
jordana.reis@vale.com

[2,3] *Departamento de Informática, Universidade Federal do Espírito Santo - UFES, Vitória, ES, 29.075-910, ES, Brazil*
rhuan.teixeira@edu.ufes.br
flavio.varejao@ufes.br

ABSTRACT

This paper proposes a deep transfer learning approach for detecting brake fluid leakage in gondola wagons using acoustic signals. Gondola wagons, also known as railroad gondolas, gondola cars, and open wagons are typically used for transporting dry cargo and rely on pneumatic brake systems that depend on compressed air components for effective braking. The study investigates three transfer-learning strategies—From-Scratch, Partial Fine-Tuning, and Full Fine-Tuning—to identify compressed air leakage based on sound emissions, mirroring the process performed by human wagon inspection professionals. Training data consists of waveform audio signals captured during real railcar inspections. In the proposed model, each audio file is processed in the time-frequency domain to obtain a mel-spectrogram, which is then used as input to pre-trained deep convolutional neural networks. Results demonstrate strong performance, achieving accuracy above 94%. The main scientific contribution lies in applying established deep learning techniques to a specific and underexplored industrial railway context, together with the development and validation of a real-world dataset collected under authentic operating conditions. These findings demonstrate the feasibility and practical effectiveness of the proposed approach for pneumatic brake leakage detection in operational environments.

1. INTRODUCTION

The railway ecosystem supports the activities of an important way of transport for many countries. According to (Pfaff, Enning, & Sutter, 2024), rail freight is by far the most energy efficient means of transportation. In 2022, railway transportation was responsible for 91% of the iron ore exported by Brazil (ANTF, Associação Nacional dos Transportadores Ferroviários, 2023). In order to supply the demand, from 1997 to 2022, the total number of wagons has increased, according to (ANTF, Associação Nacional dos Transportadores Ferroviários, 2023), from 46.816 to 112.640. As the number of these assets increases, the need for inspection grows. However, the manual process of wagon inspections has not yet reached the quality and efficiency consistent with the high-risk context inherent to wagon operational failures. In addition, it is an expensive and time-consuming activity (Hashmi et al., 2022).

One of the components that demands attention during the wagons inspection is the brake system. When transporting iron ore, the adequate wagon model to be used is the gondola type (Academic Accelerator, n.d.), which is usually sold in pairs, sharing the braking system. The gondola wagons brake system inspection includes, among other activities, the verification of fluid leakage. For this paper, the methods, data and results refer specifically to gondola wagons which use air as the brake fluid.

Figure 1 represents a wagons yard, belonging to Vale S.A., a Brazilian mining company. The train composition arrives carrying iron ore in gondola wagons, the wagons are split in batches (between 82 and 86 wagons) to be inspected.

Each of those rail vehicles accommodates around 110 tons as payload, which increases the risk of the issues resulting from an accident. Therefore, careful inspections are required in order to protect the communities, employees and the company's public image.

However, the inspection process is largely dependent on human actions. Human inspectors walk by each side of the wagon batches, at an approximate distance of 1 meter, with special attention on hearing, with the aim of listening for air leakage. This activity is performed in pairs, for the safety of both inspectors and to reduce the amount of walking per

https://doi.org/10.36001/IJPHM.2026.v17i1.4707

Figure 1. Open rail vehicle, gondola wagon.

person.

As it is always necessary to have 2 inspectors, in the eventuality of crew member absence, the activity is suspended. Since the inspection mainly uses the inspectors' hearing, it can be suspended depending on the weather, for instance, rainy days. Detecting brake fluid leaks relies heavily on auditory acuity, as small leaks are often difficult to perceive and therefore demand heightened attention, frequently leading to precautionary retention of the railcar. In contrast, larger leaks are typically identifiable from a distance because the pressurized air escaping from the compartment produces a clearly audible sound. Lastly, the adjacent rail lines must be blocked, to guarantee the inspectors safety and provide a better environment for hearing, which may impact the company productivity. The blocking of adjacent rail lines aims to mitigate the risk of workers being struck by railcars pushed along adjacent tracks, which is addressed by blocking neighboring lines on each side of the track under inspection; this most often results in two additional lines being blocked for each inspection line for a total of 3 lines being restricted from service.

This work presents a possible solution for eliminating the exposure of people to risk, by having a classification model as auxiliary tool during wagons brake fluid leakage inspections. By eliminating the need for human inspectors, there will be no suspensions due to staff shortages. Suspensions can also be reduced due to limiting conditions such as rain. Finally, there is no need to interrupt the other lines as there are no longer any threats to the safety of inspectors. Moreover, the proposed solution will contribute to increased confidence in the inspection of wagons components, reducing the risk of future unwanted events.

The proposal consists of capturing and processing audio files in the time-frequency domain to obtain the mel-spectrogram. The images generated upon this process serve as data input to pre-trained deep convolutional neural networks. Some early results were shared in a preliminary study by (Reis & Varejao, 2023). Since then, several improvements on the signal capture equipment, number of samples collected and in the classification model were implemented. The current results show that the classification models perform well, achieving 94% of accuracy, proving its ability to identify sounds of air leakage on gondola wagons brake system. Experiments also show that the proposed model performance is higher than shallow machine learning methods applied on statistical features directly extracted from the audio signal and also the deep learning models without imported pre-trained weights.

Therefore, the main contributions of this work are:

- the presentation of the problem of detecting brake fluid leaks in gondola-type wagons;
- the development of a transfer learning method for automatically detecting brake fluid leakage based on the analysis of audio signals;
- a public dataset of audio signals used in the experiments;
- a public open source experimental framework for performing brake fluid leakage detection on this audio dataset and
- an intelligent fault diagnosis method benchmark and corresponding results for the comparison of future works.
- the novel application of deep transfer learning to a limitedly explored industrial railway inspection context for brake fluid leakage detection.

The rest of this paper is organized as follows. Background knowledge and related works are presented in Section 2, the method proposal is described in Section 3, followed by Section 4 that provides the experimental methodology. Section 5 presents the results and provides analysis and discussions. Finally, the conclusion and future works are given in Section 6.

2. RELATED WORK

This section describes and overviews related work on fault detection in railway systems, on the use of acoustic signals for fault detection and on the application of transfer learning techniques in training deep models.

2.1. Railway Ecosystem Fault Detection

In the context of railway ecosystem, some studies address failures in the context of railroads in general, while few focus on the identification of fluid leakage in wagon brakes. The work of (Sakellariou, Petsounis, & Fassois, 2014) studied detecting and diagnosing failures in railway vehicle suspensions based on vibration signals. The data used was obtained from a simplified physics-based model of a railway vehicle suspension, used for vehicle simulation. (Ribeiro, Pereira, & Gama, 2016) employed a failure detection system to predict train door breakdowns. Their data was collected by sensors

installed in the doors of a Class 156 train. (Wang, Lu, Wei, & Zhang, 2019) proposed a method for bearing fault diagnosis based on frequency-domain energy feature reconstruction and composite multiscale permutation entropy. (Cheng et al., 2019) published a study on the running gears of high-speed trains, proposing a fault diagnosis method based on a semi-quantitative information model. In the context of metro trains, (Xu & Yao, 2023) studied a fault identification method for the wheelset, proposing a combination of algorithms with vibration signals as data. (Kulkarni, Qazizadeh, & Berg, 2023) presented a framework to detect rail vehicle running instability, using data gathered from sensors which serves as input for comparing anomaly detection algorithms.

Regarding the train brake system, (M. Zhang, Liu, & Dang, 2021) presented a fault diagnosis method based on multi-dimension feature fusion and Gradient Boost Decision Tree (GBTD) enhanced classification. (Liu et al., 2020) studied the health state monitoring of solenoid valves in high-speed trains, as a component of the braking system and essential factor for the safe operation of trains. The data was collected using a train brake system experimental platform, and a probabilistic neural network was applied to estimate the health status of the system. A fault diagnosis method based on multi-sensor fusion for a single fault and composite fault on train braking system was proposed by (Jin et al., 2021). (Hu, Zhang, Meng, & Kang, 2022) studied the health monitoring of high-speed train break pads, by using the vibration signal and proposing a deep subdomain generalization network to identify unhealthy pads.

Specifically related to the leakage of air brake systems in wagons, (D. Zhou, Ji, He, & Shang, 2018) presented results on the detection and isolation on the brake cylinder component, such as faults on the sensor itself, performance degradation and gas leakage. (Zuo, Ding, & Feng, 2019) diagnosed latent leakage faults of pneumatic units based on data generated from a fault simulation test platform. Statistical methods were applied to optimize the sample size selection, and the Support Vector Machine (SVM) algorithm was used for fault classification. (Xiong, Liu, & Niu, 2021) simulated leakage and stuck faults of the braking system electromagnetic valves, establishing a mathematical model and concluding that leakage fault could be diagnosed using the time-domain analysis of the pressure signal on the mathematical model built. (Zanelli et al., 2022) presented a framework for implementing real-time wireless sensor monitoring of pressure variation inside the braking system of a freight wagon. For optimizing the monitoring of vibrations on dynamometric test rigs for railway brakes, (Pugi, Rosano, Viviani, Cabrucci, & Bocciolini, 2023) proposed to apply simple finite element models to monitor vibration levels and rotor-dynamical behaviour by using accelerometers to measure the vibrations of bearing. (Ge, Chen, Ling, Zhai, & Wang, 2023) proposed an approach to realize modeling and solving the air brake system based on fluid dynamics theory, with a simulation model development for locomotives and another for wagons.

To the best of our knowledge no work has ever used acoustic signals to detect brake fluid leakage in any type of wagons.

2.2. Acoustic Signal Fault Detection

Acoustic signal classification is a wide area of research ranging from speech recognition to song identification. This subsection aims to review works that used acoustic signals for fault detection or sound classification. Particular emphasis is given to studies that used the melspectrogram, an acoustic signal processing technique.

The sound emission of a gear-set is used to evaluate a fault diagnose technique proposed by (Wu & Chan, 2009), applying continuous wavelet transform combined with energy spectrum feature selection. Their experiments were performed on a test platform, with a condenser microphone located beside the gear-set platform to measure the sound emission signal. The authors presented a neural network model with 98% accuracy in recognizing faults. Sound signal analysis was presented as an approach to detect and localize porosity in gas pipelines steel by (Yusof, Kamaruzaman, Ishak, & Ghazali, 2017). Sound capturing was performed, during a welding process, by a microphone with bandwidth of 20 to 10,000 Hz, attached to the welding torch to ensure the distance as a constant. (Sangeetha & Hemamalini, 2017) presented a study to monitor the condition of induction machines by predicting the torque from acoustic signals. The sound was acquired by a microphone placed 1 cm away from the machine, and analyzed using a dyadic wavelet transform, which served as input for a Pseudo spectrum Multiple Signal Classification algorithm. (Yao et al., 2018) proposed a deep learning gear fault diagnosis based on sound signal analysis, getting the audio files generated by acoustic sensors installed around the gears and applying Fourier transform to each sample before using them as data set for a Convolutional Neural Network (CNN). Environmental Sound Classification (ESC) has been the object of the study published by (Y. Chen, Guo, Liang, Wang, & Qian, 2019), who applied dilated convolutional neural network to classify urban sounds. (Ye, Zhang, & Liang, 2020) used acoustic signals, by capturing the sound of the train wheels during the uncoupling process on the hump platform. According to their study, the sound from the wheels becomes different when the brake is not released, which was the input data set for a hybrid algorithm framework proposed by the authors. For diagnosing faults in worn gearboxes, (Karabacak, Özmen, & Gümüşel, 2022) combined data from vibration, sound and thermal image information to be used as input for Adversarial Neural Network (ANN) and SVM algorithms. The microphone used to capture the sound was placed 100 mm from the gearboxes, and the samples were transformed by applying FFT. Acoustic signal processing is classified as data analysis

in time-domain, and waveform analysis in frequency and time-frequency domain, as noted by (Sangeetha & Hemamalini, 2017). According to (Giorgi, Levy, & Apple, 2022), the mel-spectrogram is a low-resolution time-frequency representation derived from the power spectrogram. Additionally, (Umesh, Cohen, & Nelson, 1999) states that the Mel scale is a fundamental result of psychoacoustics, relating real frequency with perceived frequency, which is the frequency the human ear can recognize or perceive. (Luz, Oliveira, Araújo, & Magalhães, 2021) applied mel-spectrograms as time-frequency representations to provide data to a CNN model aiming to classify urban sounds. After the signal partition, it was computed the Discrete Fourier over the overlapping windows and employed Librosa (McFee et al., 2015) to generate the mel-spectrograms. (Q. Zhou et al., 2021) used the audio signal as input for a cough classification model. They audio-segmented the file, extracted the features applying mel-spectrogram, normalized the results and finally, used them as dataset for a CNN-based model. (Ustubioglu, Ustubioglu, & Ulutas, 2023) utilized Short-Time Fourier Transform (STFT) to create the spectrogram of the speech signal, previously split in 30 ms frames and each multiplied by the Hamming window, targeting to detect audio copy-move forgery. The spectrum of the sound, represented as an image, was used in the work of (Tagawa, Maskeliūnas, & Damaševičius, 2021), where each segment of waveform sound data is processed in FFT and then applied the mel-spectrogram. The images are used for anomaly detection of mechanical failures.

2.3. Transfer Learning

Transfer learning is the reuse of a pre-trained model on a new problem. This technique is popular in deep learning because it can train deep neural networks with less data. As most real-world problems typically do not have millions of labelled data, this technique becomes useful for training such complex models.

(Rodrigues, Jutten, & Congedo, 2023) presented a Transfer Learning approach to handle the statistical variability of eletroencephalographic (EEG) signals recorded on different sessions and/or different subjects. The Transfer Learning contribution was based on geometrical transformations, such as translation, rotation and scaling. A clustering-guided balanced domain adaptation transfer learning is proposed by (T. Zhang, Peng, Tang, Yan, & Deng, 2023) to solve the difference in the labelling method and the noise for simulation domain and measurement datasets in the context of robots error prediction.

According to (Z. Chen et al., 2023), Deep Transfer Learning (DTL) leverages knowledge or extensive data sets from related domains to enhance the prediction accuracy and generalization performance of models by transferring the information to similar target tasks. (Padha & Sahoo, 2023) employed transfer learning to improve the ability of Quantum Long Short-Term Memory (LSTM) model further, due to the presence of small labelled data set. In a similar direction (Ma et al., 2024) highlighted that DTL has emerged as a powerful technique and offers a promising solution for Thermal EM with scarce labelled samples. In the context of in-vehicle intrusion detection system, (Hoang & Kim, 2022) also made use of transfer learning technique to improve the performance of a limited-size data set and over-fitting issues avoidance.

No studies were found that directly address acoustic emission sensors for detecting air leaks in railway braking systems. However, this may be a potential avenue for future research.

3. LEAKAGE DETECTION BASED ON TRANSFER LEARNING

Due to the limiting conditions of capturing brake fluid leakage recordings described in Section 1, this study counts on a small set of audio file samples. To address this situation the present study employs transfer learning. Pre-trained networks were executed with both importing and not importing weights.

Sample sufficiency was determined empirically when transfer learning–based models began to exhibit consistent results, evidenced by reduced variability in performance metrics and increased stability during training. These indicators suggested that the available dataset provided adequate support for model generalization, therefore deemed the sample size to be sufficient based on the observed stability and reproducibility of the results.

A pre-trained model refers to a saved network that has been previously trained on a substantial dataset, often in the context of a large-scale image-classification task. Users have the option to either directly utilize the pre-trained model as it is or tailor it to a specific task by customizing its architecture or parameters.

This work proposes a method for identifying fluid leakage in gondola wagon brakes using deep transfer learning, leveraging images derived from transformed acoustic signals as the input dataset. The process involves capturing acoustic sound signals during routine inspections of wagon brake systems in waveform audio file format and subsequently transforming them into 2D images. This approach, previously applied in fault detection domains, such as monitoring the health of milling cutter tools by transforming vibration data into spectrograms (Patil, Pardeshi, & Patange, 2023), is adopted in this study. Since the primary emphasis is on identifying leakage sounds discernible to the human ear, instead of using the conventional spectrogram, this work utilizes the Mel scale spectrogram to produce 2D image representations for the acoustic signals.

The subsequent subsections outline the overarching process proposed in this work and provide a succinct overview of key aspects related to the melspectrogram and transfer learning techniques.

3.1. Overview of the Transfer Learning Approach

Figure 2 illustrates the workflow used in this work. It initiates in the wagons yard with the process of sampling the wagon brake system during inspections, capturing waveform audio files. Utilizing acoustic signals for data analysis enables the exploration of time-frequency domain characteristics in the sound recordings, facilitating pattern recognition in leakage samples. The audio capturing process yields a dataset of waveform files, recorded in the WAV format. Waveforms offer a direct representation of the amplitude variations of an audio signal over time. This amplitude-time depiction encapsulates essential information regarding the sound's intensity and duration.

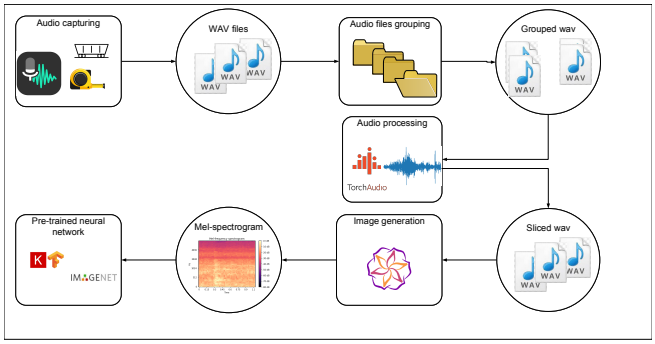

Figure 2. Method overview. The squares represent processes and the circles represent data generated by the processes. After capturing the sounds, files are segmented and transformed to 2D images. The images are used for training or fine-tuning the neural networks.

The next workflow process segregates the samples into groups to avoid the similarity bias (Rauber, Loca, Boldt, Rodrigues, & Varejão, 2021) that occurs when multiple patterns are extracted from a single signal acquisition, resulting in patterns that are very similar in both training and testing datasets. The proper separation of samples in different groups ensure that the training data used for model development is significantly different from the samples used for testing the model.

The outcome of the grouping process is a balanced dataset for each group. The balancing process was conducted based on sample classes, ensuring an equal number of recordings were collected for both normal and failure classes.

The subsequent step involves audio processing, which includes segmenting the WAV files into multiple smaller samples, thereby creating a larger dataset. Each audio segment has a fixed duration, set to 1 second for this study. This time interval size was empirically checked and shown sufficient for the method proposed in this work. The segmentation enhances feature extraction and allows for the individual treatment of segments based on their spectral characteristics. Importantly, each sample segment remains within the same dataset analysis group, whether it is for training, testing, or validation purposes. Indeed, each group becomes a fold further used to cross-validation experiments.

The next workflow process consists of converting the acoustic signals to images. Acoustic signals can be effectively represented as images by transforming them into spectrograms. Considering that the brake system leakage in gondola wagons are detected by the sound listened by human inspectors, this work transforms the waveform files to Mel-spectrograms, which are spectrograms created in the Mel scale. This allows for representing the sound in a way closer to what the human auditory system recognizes.

The final processing step of the workflow uses the Mel-spectrogram images to train CNNs from the scratch or fine tune pre-trained deep neural networks provenient from the *ImageNet* (Deng et al., 2009) dataset to the task of detection of brake system leakage in gondola wagons. The fine-tuning process allows the network to adapt its learned features to the specific characteristics of the Mel-spectrogram dataset, optimizing its performance for the task at hand. This work experimented with four architectures of pre-trained CNNs. They were used both for training from the scratch and for fine-tuning pre-trained networks.

The next two subsections present in more details the audio signal to image transformation using Mel-spectrograms and the training/fine tuning of deep neural networks employed in this work.

3.2. Mel-spectrograms

According to (S. Chen, 2022), a spectrogram is a representation that displays signal strength over time at various frequencies. It can be visualized in two-dimensional graphs with the color as the third variable, indicating the power of an amplitude at a particular time through color intensity or brightness. Essentially, a spectrogram provides a visual representation of a sound's frequency spectrum as a function of time (S. Chen, 2022).

At a high-level, spectrograms are obtained by splitting the original signal into overlapping short-time frames of equal length. For each resulting frame, a window function, specifically the Blackmann-Harris function in this work, is applied to control spectral leakage to adjacent lobes. Fourier transform is then applied to each short-frame, revealing the Fourier spectrum for those short-frames. The Short-Time Fourier Transform was chosen for this study due to its ability to uncover details and patterns by breaking the signal and analyzing its frequency individually, which might be lost with traditional Fourier transform.

Mathematically, for a given signal x(t), the complex signal for the frequency band k, in the time t, can be represented as

follows:

$$X_t(\mathrm{e}^{jw_k}) = \sum_m x(m)w(t-n)\mathrm{e}^{-jw_k m}k = 0, 1, ..., n-1 \tag{1}$$

where

$$w_k = (2\pi k)/n \tag{2}$$

is the frequency in radians, n is the number of frequency bands, m represents the discrete time index for the signal x, j is the imaginary unit and $w(m)$ is the selected window function.

The power spectrum density is obtained by squaring the magnitude of the frequency spectrum for each frame. The spectrogram is then created by transfering the power into the decibel scale, commonly used to measure sound intensity. Figure 3 provides a representation of a simple spectrogram, depicting an acoustic signal of air brake leakage captured during the data collection process for this study.

The mel-spectrogram is a conversion of the frequency (Hz) into Mel scale by applying Mel filter banks to the spectrogram. This is performed to represent the sound in a way closer to what the human auditory system can recognize. The Mel filter bank outputs a frequency-domain, decomposing an audio signal into separate frequency bands in the Mel frequency scale, which mimics the non-linear human perception of sound.

The mel-spectrogram can be mathematically represented as the log-scaled energy of short-time spectral components projected onto a mel-spaced filter bank. Starting from an audio signal $x[n]$, its short-time Fourier transform $X(m,k)$ is computed, the power spectrum is obtained as $|X(m,k)|^2$, and each time frame is mapped to mel bands through triangular filters $M_r(k)$. The resulting representation is

$$S_{\mathrm{mel}}(m,r) = \sum_k |X(m,k)|^2 M_r(k) \tag{3}$$

and logarithmic compression gives

$$\widetilde{S}_{\mathrm{mel}}(m,r) = \log\big(S_{\mathrm{mel}}(m,r) + \epsilon\big) \tag{4}$$

where mel filter centers are defined from

$$\mathrm{mel}(f) = 2595\log_{10}(1 + f/700) \tag{5}$$

The same fragment of sound, with the duration of 1 second, represented in Figure 3 as a spectrogram is presented in figure 4 as a mel-spectrogram.

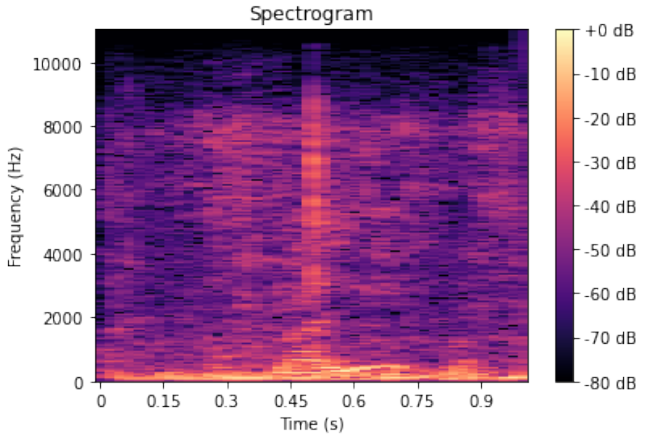

Figure 3. Spectrogram Representation. Image corresponding to 1 second duration of an acoustic signal of air brake leakage.

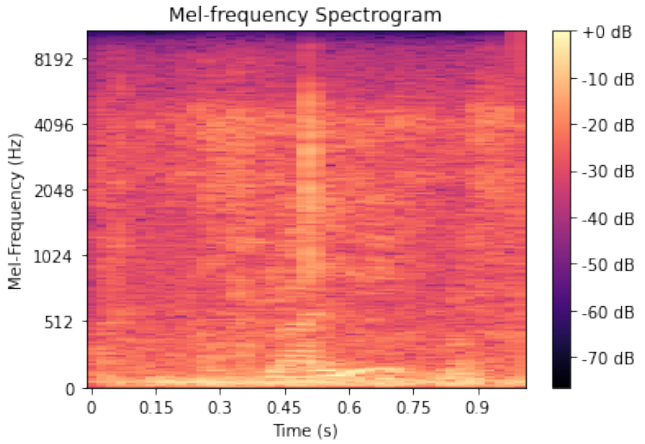

Figure 4. Mel-spectrogram Representation. Image corresponding to the acoustic signal of Figure 3 in mel-scale.

Compared to the spectrogram, the mel-spectrogram provides a visualization of variations in frequencies, and the characteristics of power distribution are clearly distinct in both time and frequency domains.

3.3. Pre-trained Deep Networks

Convolutional Neural Networks (CNN) are mainly used for images and speech recognition due to its convolutional layer capability of reducing the high dimensionality of those images without information loss. CNNs, especially in computer vision tasks, are adept at detecting edges in earlier layers, shapes in middle layers, and some task-specific features in later layers.

Given that the acoustic signal can be represented as an image, CNNs are a suitable approach for recognizing patterns within these images. Considering the small dataset collected, reusing a pre-trained model, expedites the insights gained from the

proposed method, yielding results in a shorter period of time, comparing to the complete process of building and training of a deep neural network. Keras (Wood et al., 2022) provides lots of pre-trained models suitable for Transfer Learning, including some specifically trained on the *Imagenet* (Deng et al., 2009) datasets. *Imagenet* is a substantial collection of images organized within the framework of the WordNet structure, providing a comprehensive ontology. The neural networks presented in this study are pre-trained using images from this database.

There is an optional step to the transfer learning, the fine-tuning, which consists of unfreezing the entire model obtained, or part of it, and re-training it on the new data with a very low learning rate. This enables the model to adapt and learn more specific features related to the new task. A partial fine-tuning comprehends the freezing of all the pre-trained layers, training only the new layers added on top. In a full fine-tuning, some of the pre-trained layers are unfrozen and then, allowed to be updated during training. (Vrbančič & Podgorelec, 2020) explain that currently, there is no general rule or recipe to follow in order to determine which layers to fine-tune or which hyper-parameter settings to use. Most of the decisions are based on previous experiences of dealing with such problems.

Following a similar approach employed by (Kensert, Harrison, & Spjuth, 2018), who applied ResNet50, InceptionV3 and InceptionResNetV2 to predict cellular morphological changes based on images, the experiment results presented in this paper regards to the subsequent neural networks: Xception, ResNet50V2, ResNet152V2 and InceptionResNetV2. These CNNs were chosen based on their availability in newer versions and estimated top 5 accuracy informed by *Keras* (Wood et al., 2022) documentation.

The ResNet and Inception families are among the most popular CNN applied to image classification, also leveraged by an annual context called Imagenet Large Scale Visual Recognition Challenge (ILSVRC). Xception was developed by *Google* researches as an interpretation of InceptionV3 and presents good performance while not requiring large computing resources.

4. EXPERIMENTAL METHODOLOGY

This section describes the dataset used in the experiments, the evaluation metrics and statistical tests used for comparing the fault diagnosis classifiers performance, the resampling strategy used for defining training, validation and testing subsets, and finally the shallow and deep classifiers evaluated in this study.

4.1. Dataset

The results of this work are based on a dataset comprising a total of 84 recordings instances collected between December 2022 and May 2023. The dataset is intentionally balanced, consisting of 42 instances of *failure* conditions and 42 instances of *normal* conditions. Table 1 reports the number of recordings collected per inspection line on each date, providing a detailed view of the dataset composition and the coverage of different operational conditions. In total, 84 recordings were collected along 5 inspection days and 14 rail lines inspected. The recordings were obtained during routine inspections of wagon brakes at a yard owned by Vale S. A., a mining company with operations in Brazil, with part of its product transportation conveyed by cargo trains.

The approximate number of railcars is considered equivalent to the number of recordings performed, totaling 84, under the assumption that each recording represents a single railcar inspection. According to the adopted procedure, any railcar with a brake fluid leak was retained for corrective maintenance; however, inspections were conducted over a one-year period without longitudinal tracking. Therefore, it is possible that the same railcar was recorded in multiple inspections during the study period.

The sound acquisition setup was designed to ensure adequate acoustic quality and experimental reproducibility using accessible, low-cost resources. Compressed air leak recordings were performed with a smartphone and the WaveEditor application in WAV format, maintaining the original waveform for processing. A standard distance of 100 cm from the sound source was used, adjusted when necessary for terrain or safety considerations, and a unidirectional cardioid microphone minimized environmental noise while capturing signals within the 30 Hz to 15 kHz range. This configuration optimized signal quality, reduced interference, and ensured methodological consistency in data acquisition.

Table 1. Distribution of failure and healthy recordings by inspection date and line.

Date	Line	Failure Samples	Healthy Samples
2022-12-16	03	1	0
	05	5	6
	26	7	0
2023-03-03	01	0	4
	14	1	7
	20	2	1
	21	3	0
	23	3	0
2023-03-24	02	2	0
	20	3	0
2023-04-14	10	3	0
2023-04-28	06	3	2
2023-05-05	01	1	13
	20	3	5
	28	1	2
2023-06-02	01	0	1
	17	4	1
Total	—	**42**	**42**

It is important to note that, to capture the recording instances, certain restrictions were necessary to adhere to health and

safety regulations. The recordings were taken while walking through only one side of the wagon batches. No inspections were conducted on rainy days or in the absence of one of the inspectors. The recording process involved the use of a measuring tape to ensure that the device's distance from the sound source would be approximately the same as the inspector distance during an inspection routine. Each recording lasted approximately 5 seconds and was labelled with healthy or failure, as well as the inspection line where it was taken. Few recordings may slightly exceed or fall short of the 5-second duration due to the inherent start-stop nature of the recording process, an expected behaviour.

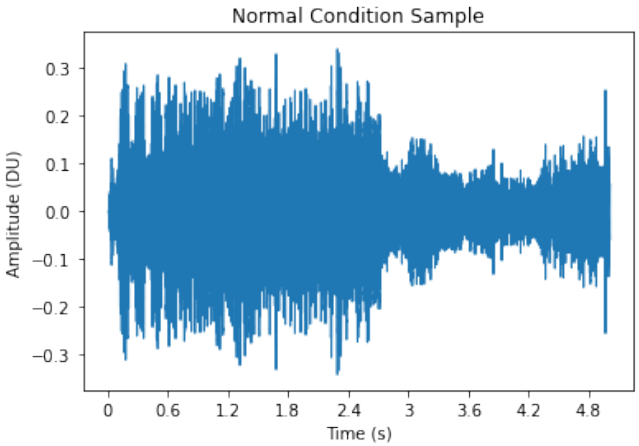

Figure 5. An illustration of an audio sample representing normal conditions. The signal representation varies among the samples, indicating that not all of them follow the same pattern.

Figure 5 illustrates a waveform representing a normal condition instance, whereas Figure 6 presents a similar representation for a failure condition instance. It is important to highlight that there is heterogeneity in signal representation among the samples, even those belonging to the same class. Therefore, Figures 5 and 6, both lasting 5 seconds of duration, serve as illustrations and do not fully capture the signal variability present in all samples labelled as normal or failure conditions. The amplitude measurement unit is represented by the abstraction of Digital Unit (DU), since those samples are digitally captured.

Given the less frequent natural occurrence of failure events compared to normal conditions, this study intentionally balanced the samples. For every failure condition audio recording, a corresponding audio recording of a normal condition was included, both taken at the wagons yard.

Aiming to increase the number of samples, each recording file was segmented in smaller files lasting 1 second at maximum. As a result, at least 4 files lasting 1 second were created per wav file. As the original recordings' duration lasts approximately 5 seconds, there are samples presenting only 4 segmented files lasting 1 second and the 5th lasting less than 5 seconds. The same situation happens with a few samples lasting more than 5 seconds, which were segmented in 5 files lasting 1 second and the 6th lasting less than 1 second. The smaller than 5 seconds segments were discarded. This process generated a total of 491 audio samples per round, where 251 was labelled as normal condition and 240 are samples representing failure condition.

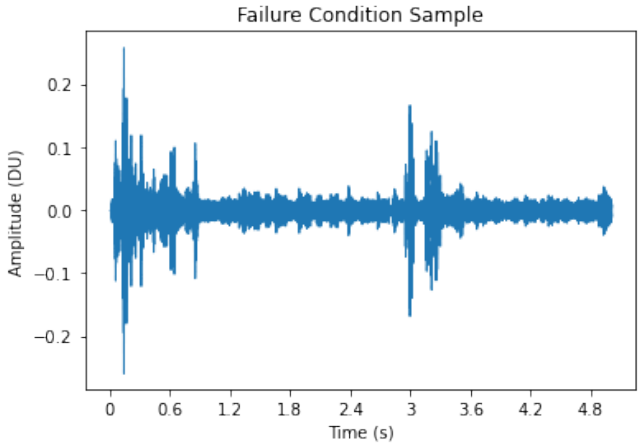

Figure 6. An illustration of audio sample for leakage (failure) condition. The signal representation varies among the samples, indicating that not all of them follow the same pattern.

Dataset independence is defined at the recording-file level. Accordingly, samples are grouped prior to any processing or segmentation so that all segments originating from the same recording remain within the same subset, preventing information leakage between the training, validation, and test sets. This strategy preserves statistical independence among datasets and supports a reliable evaluation of model generalization performance.

4.2. Performance Evaluation Measures

Accuracy, a widely-used metric for evaluating classification models, is employed in the evaluation of the models experimented in this work. The accuracy of a model is computed by dividing the number of correct predictions by the total number of predictions. For binary classification and balanced data sets, accuracy is generally considered a reliable metric for performance evaluation (Godbole, Dahl, Gilmer, Shallue, & Nado, 2022).

The Area Under the ROC Curve (AUC) is utilized as another valuable metric for assessing the performance of binary classification models. AUC is an aggregate measurement across possible classification thresholds, representing the probability that the model ranks a random positive sample more highly than a random negative sample (Godbole et al., 2022).

This study provides both metrics for a comprehensive analysis of the models' performance. It is important to emphasize that the Area Under the ROC Curve (AUC) is inherently suitable for evaluating model performance in the presence of class imbalance, which may be advantageous for future analyses. The calculations for both metrics were conducted using the functions available in *Scikit-Learn* (Pedregosa et al., 2011).

The performance of the models is assessed through an analysis that involves comparing the accuracy and AUC metrics generated during their execution. This comparison is conducted using paired-samples *t-student* and non-parametric *Wilcoxon Signed-Rank* statistical tests. The t-student test gives the means between each pair of models results for both metrics aforementioned whereas the Wilcoxon test gives the sample location comparison between each pair of models results for both metrics. In summary, this study analyses how similarly the models are performing from the point of view of general average, and also from the point of view of ranking or performance range.

4.3. Repeated Stratified Nested Group K-fold Cross Validation

Repeated stratified nested cross-validation process for hyperparameter value selection and model evaluation is used as resampling strategy. Nested cross-validation is recommended when multiple tuning parameters are estimated, generating multiple layers of cross-validation loops. This approach reduces the risk of overfitting to the dataset when performing the search procedure. The outer loop was set as k = 10 and the inner loop was set as k = 3.

In order to avoid the similarity bias problem (Rauber et al., 2021), distinct samples from the same recording instance should not be present in both the training and testing sets. Thus, samples segmented from each recording instance are exclusively assigned to one of the 10 folds of the outer loop. The Group k-fold method is used with this purpose.

The t-test requires observations to be normally distributed. The central limit theorem (CLT) asserts that as the sample size increases, the distribution of sample means approximates normality, irrespective of the underlying population distribution. Typically, sample sizes equal to or exceeding 30 are deemed adequate for the CLT to apply. Therefore, three rounds of nested cross validation are performed, each one with a different division of recording instances in the 10 folds of the outer loop. This strategy produces 30 different testing evaluations enabling the reliable application of the t-test.

4.4. Statistical Features and Shallow Machine Learning Classifiers

Evaluating the performance of a proposed classification model involves comparing the results obtained by the model with those of baseline classification models. In this work, the baselines use statistical features directly extracted from the audio signals samples and classifiers based on techniques of shallow machine learning. The statistical features use the ones presented in (Arunkumar & Manjunath, 2016). Mean and Median were also included in the extracted features set. Each feature is mathematically defined as presented in Table 2, where i is the sample identification, N is the total number of samples, and $X(i)$ is the signal value for a given sample i:

The shallow machine learning classification models used in this work are:

Table 3. Classifiers' hyperparameters values tuned by grid search.

Classifier	Hyperparameter	Values
RFC	Number of trees	10, 100 and 1000
XGBoost	Number of trees	10, 100 and 1000
XGBoost	Maximum depth	2 and 6
XGBoost	Learning rate	0.1 and 0.01
LightGBM	Number of leaves	31 and 127
LightGBM	Reg_alpha (L1)	0.1 and 0.5
LightGBM	Minimum data in a leaf	30, 50, 100, 300 and 400
LightGBM	Lambda_l1	0, 1 and 1.5
LightGBM	Lambda_l2	0 and 1

4.5. Transfer Learning Methods

This work employed four neural networks available for transfer learning in *Keras* (Wood et al., 2022): Xception, ResNet50V2, ResNet152V2 and InceptionResNetV2. Three distinct transfer learning approaches were investigated. The original top layer of the pre-trained model was removed to adapt the architecture to the specific requirements of the current task, as this layer is typically tailored to a different set of classes and objectives. It was replaced with a 2D global average pooling layer followed by a dense layer with sigmoid activation for binary classification, preserving the previously learned deep representations while enabling refinement of the decision-making stage. This modification improves the model's ability to specialize for the proposed task and enhances alignment between the network architecture and the dataset characteristics.

The first approach uses partial fine-tuning (PFT). Beyond the new output layer, the five latter layers of the networks were also unfrozen to be updated during training. The second approach conducted a full, or complete, fine tuning (CFT), by fully fine tuning the weights of all layers of the pre-trained model. Thus, the weights of all layers were unfrozen to be updated during training. The third approach conducted a training from scratch (FST), by applying the pre-trained model without any weight importing. Therefore, only the networks topology is transferred in this approach.

The following statements hold true for all approaches. Each neural network model was configured to run 100 epochs. The

Table 2. Statistical features extracted from the waveform signal.

Characteristic	Mathematical definition	Description
Mean	$\overline{X} = \frac{\sum_{i=1}^{N} X(i)}{N}$	Represents the mathematical average of a given dataset.
Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[X(i) - \overline{X}]^2}$	Represents the amount of dispersion or variation from the average.
Root mean square value	$\text{RMSV} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[X(i)]^2}$	Indicates the amount of energy in the signal.
Peak value (PV)	$\text{PV} = \max X(i)$	Indicates the largest amplitude value of the signal.
Crest factor (CV)	$\text{CV} = \frac{PV}{RMSV}$	Defined as the ratio of the PV to the RMSV, indicating the existence of sharp peaks in signal.
Kurtosis	$\text{Kurtosis} = \frac{1}{(N-1)\sigma^4}\sum_{i=1}^{N}[X(i) - \overline{X}]^4$	Indicates the property of the signal, peaked or flat, relative to a normal distribution of the signal.
Skewness	$\text{Skewness} = \frac{1}{(N-1)\sigma^3}\sum_{i=1}^{N}[X(i) - \overline{X}]^3$	Measures symmetry in a distribution and indicates the position and orientation of a defect.
Clearance factor (CF)	$\text{CF} = \frac{PV}{(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert X(i)\rvert})^2}$	Ratio of the signal's PV to the square of the average of the absolute signal's square root value.
Impulse factor (IF)	$\text{IF} = \frac{PV}{\frac{1}{N}\sum_{i=1}^{N}\lvert X(i)\rvert}$	The ratio of the signal's PV to the average absolute signal value.
Shape factor (SF)	$\text{SF} = \frac{RMSV}{\frac{1}{N}\sum_{i=1}^{N}\lvert X(i)\rvert}$	The ratio of the signal's RMSV to the average absolute signal value.
Histogram upper bound	$\text{upper bound} = \max X_i + \frac{0.5[\max X_i - \min X_i]}{N-1}$	Measures the highest value within the range of the data being analyzed.
Histogram lower bound	$\text{lower bound} = \min X_i - \frac{0.5[\max X_i - \min X_i]}{N-1}$	Measures the lowest value within the range of the data being analyzed.
Median	If N is even: $\tilde{x} = \frac{\left(\frac{N}{2}\right)^{\text{th}}\text{ obs.} + \left(\frac{N+1}{2}\right)^{\text{th}}\text{ obs.}}{2}$ If N is odd: $\tilde{x} = \left(\frac{N+1}{2}\right)^{\text{th}}\text{ obs.}$	Represents the midpoint value of a given dataset.

Adam optimizer function was used for the model compilation. Binary cross-entropy was set as the *loss* parameter. The batch size for image samples training was set as 10, while the number of batches to be yielded per epoch was set as 5. An early stopping callback was set to monitor the loss function values, and restore the best weights if there is no improvement in performance with loss decreasing after 5 epochs. The input shape of the data provided to the model was 100x100x3 and color mode as *RGB*.

Regarding to the models' input shape, images generated from the acoustic signals were used. As deep-learning models usually exhibit good performance when presented with images as input, the Librosa (McFee et al., 2015) library was employed to produce the signals' mel-spectrograms, by applying the Short Time Fourier Transformation (STFT). The spectrogram was configured with the following parameters values. The *frame length* was set as 2048, the *hop length* as 1024, and the *window* as the result of the Blackmann-Harris Librosa function execution for 1024 number of points in the output window. Finally, for each waveform file, a mel-spectrogram image was generated using the Librosa functions. The process involved setting the sample rate to 44.100 and using 100 bands of frequency.

5. RESULTS

This section presents the experimental results obtained in this study. Classifiers models utilizing statistical features were tested to establish a baseline for comparison. Table 4 illustrates the performance of the statistical models in terms of average accuracy (ACC) and average area under the ROC curve (AUC) of the 30 test evaluations. Standard deviation values (σ) are presented in parenthesis. The items in this table are arranged in descending order for both ACC and AUC.

Table 4. Statistical models performance on the available dataset.

Model	ACC (σ)	AUC (σ)
RFC	79.52 (10.83)	85.23 (11.42)
LightGBM	77.73 (10.83)	85.07 (10.98)
XGBoost	77.48 (11.65)	84.32 (11.41)

As evidenced by the results, the models performed relatively well, with an average ACC above 77% and an average AUC above 84%. The best baseline model performance was the random forest (RFC) yielding a 79.52% value of ACC and a 85.25% value of AUC.

The results of the experiments using the transfer learning approaches are presented in Table 5. One additional column is added to the table. It indicates the transfer learning approach used by the respective method. The items in this table are displayed from the highest to the lowest ACC values.

The superiority of the transfer approaches with fine tuning, both PFT and CFT, is evident in both metrics - ACC and AUC. Results are better than the FST and baseline models. Furthermore, the standard deviation in these metrics is also lower than those of the other models.

Analyzing the performances of the deep transfer learning models, one may note that those trained from scratch did not

Table 5. Deep transfer learning models performance on the available dataset.

Model	Transfer Learning Approach	ACC (σ)	AUC (σ)
Xception	PFT	96.98 (8.47)	98.17 (6.13)
ResNet50V2	PFT	96.12 (9.60)	96.91 (8.68)
ResNet50V2	CFT	95.91 (9.36)	97.61 (6.61)
Xception	CFT	95.25 (8.41)	97.47 (6.28)
InceptionResNetV2	CFT	94.84 (8.40)	96.60 (7.51)
ResNet152V2	PFT	94.48 (10.74)	96.99 (7.81)
InceptionResNetV2	PFT	94.30 (8.57)	96.66 (7.26)
ResNet152V2	CFT	93.67 (11.07)	96.25 (8.35)
Xception	FST	87.51 (12.79)	95.08 (9.92)
ResNet50V2	FST	76.77 (15.27)	93.28 (12.14)
ResNet152V2	FST	74.16 (22.25)	88.33 (15.92)
InceptionResNetV2	FST	68.92 (15.27)	81.31 (15.77)

perform better than the statistical models reported in Table 4. This was expected due to the small amount of samples in the dataset. On the other hand, when compared to the models which imported the pre-trained weights, there is a significant difference in performance.

The deep transfer learning models showed similar performances when trained with partial or full fine tuning, whereas the two best models use the partial approach. Besides, the best result (96.98% of ACC and 98.17% of AUC) was achieved with the PFT approach with the Xception network. This model also had the lowest standard deviation (8.47% of ACC and 6.13% of AUC), indicating better regularity. These results suggest that partial tuning is a better approach, as it achieves high performance while keeping computational costs lower than full tuning training.

To further investigate the significance of the observed performance differences, statistical analysis employing the student's t-test and the Wilcoxon test were performed. The comparison is performed among the ACC and AUC values obtained from the experimental results of the deep transfer learning models trained from scratch, those with imported pre-trained weights, both partially and fully fine-tuned, and the statistical models. The significance level for both tests will be considered as 0.05, meaning that values below this threshold will indicate a rejection of the null hypothesis, concluding that there is a statistically significant difference in performance between the two populations of models results.

Table 6 presents a pairwise comparison of the 15 models' ACC performance using t student and Wilcoxon p-values. Wilcoxon p-values are represented in the lower triangle of the table, while t student are represented in the upper triangle.

One discernible pattern emerges upon examination of Table 6. The fine-tuning transfer learning models (PFT and CFT) exhibit statistically distinct performance than the statistical models (RFC, LightGBM, and XGB) and the from scratch transfer learning models (FST). Unless the PFT Xception model, there is almost no statistical difference between the PFT and CFT models. The same observation may be made between the statistical models and the transfer learning models trained from scratch. This observation substantiates the findings outlined in the analyses conducted in Tables 4 and 5.

To further enhance the analysis, Table 7 provides a parallel examination of the 15 distinct models, this time focusing on the AUC score. Basically, results are similar to those of Table 6, and the same observations are valid.

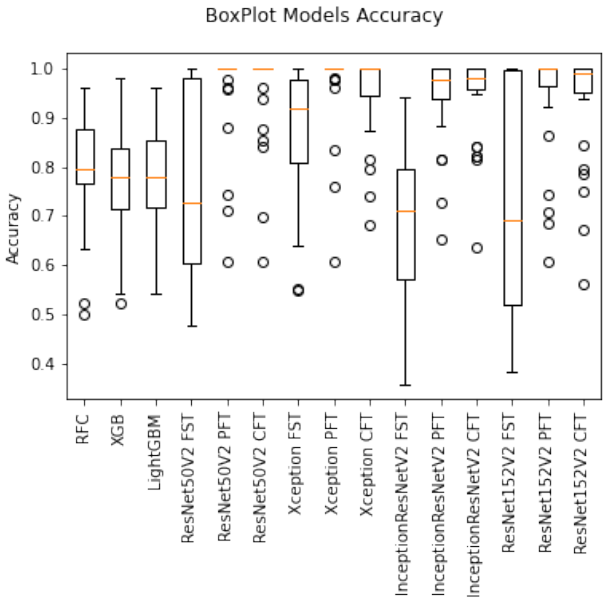

Figure 7. Boxplot of the models ACC performance.

The boxplots of figures 7 and 8 provide insights into the spread and central tendency of ACC and AUC scores distribution of all evaluated models.

In Figure 7 the central tendency of the data distribution, indicated by the middle line, reveals that the median accuracy for both statistical models and FST models falls between 60% and 80%, except for Xception FST, which surpasses 90%. Many models exhibit outliers, signaling potential anomalies probably related to the dataset limitation, demanding further investigation. PFT and CFT models demonstrate consistent

Table 6. ACC Hypothesis Tests Results; p-value pair comparison using t-student and Wilcoxon tests. t-student p-values are displayed in the upper triangle, while Wilcoxon values in the lower triangle. Statistically differences in the results at a 0.05 significance level are highlighted in orange color.

T-student / Wilcoxon	RFC	LightGBM	XGB	InceptionResNetV2 (FST)	ResNet152V2 (FST)	Xception (FST)	ResNet50V2 (FST)	InceptionResNetV2 (PFT)	ResNet152V2 (PFT)	Xception (PFT)	ResNet50V2 (PFT)	InceptionResNetV2 (CFT)	ResNet152V2 (CFT)	Xception (CFT)	ResNet50V2 (CFT)
RFC		0.09	0.09	0.00	0.27	0.02	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LightGBM	0.12		0.81	0.01	0.43	0.00	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
XGB	0.09	0.85		0.01	0.49	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InceptionResNetV2 (FST)	0.00	0.01	0.01		0.24	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ResNet152V2 (FST)	0.32	0.50	0.55	0.27		0.01	0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Xception (FST)	0.02	0.01	0.01	0.00	0.01		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ResNet50V2 (FST)	0.47	0.75	0.81	0.02	0.57	0.01		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InceptionResNetV2 (PFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.87	0.00	0.11	0.51	0.45	0.23	0.01
ResNet152V2 (PFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69		0.02	0.22	0.77	0.50	0.57	0.19
Xception (PFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01		0.43	0.03	0.01	0.14	0.19
ResNet50V2 (PFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.12	0.50		0.18	0.08	0.42	0.85
InceptionResNetV2 (CFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	1.00	0.03	0.13		0.22	0.50	0.21
ResNet152V2 (CFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.15	0.00	0.02	0.23		0.05	0.00
Xception (CFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.97	0.12	0.44	0.37	0.05		0.35
ResNet50V2 (CFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.24	0.24	0.35	0.11	0.00	0.42	

Table 7. AUC Hypothesis Tests Results; p-value pair comparison using t-student and Wilcoxon tests. T-student p-values are displayed in the upper triangle, while Wilcoxon values in the lower triangle. Statistically significant differences (p ¡ 0.05) are highlighted in orange.

T-student / Wilcoxon	RFC	LightGBM	XGB	InceptionResNetV2 (FST)	ResNet152V2 (FST)	Xception (FST)	ResNet50V2 (FST)	InceptionResNetV2 (PFT)	ResNet152V2 (PFT)	Xception (PFT)	ResNet50V2 (PFT)	InceptionResNetV2 (CFT)	ResNet152V2 (CFT)	Xception (CFT)	ResNet50V2 (CFT)
RFC		0.83	0.32	0.15	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LightGBM	0.85		0.46	0.21	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
XGB	0.74	0.65		0.33	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InceptionResNetV2 (FST)	0.25	0.38	0.28		0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ResNet152V2 (FST)	0.15	0.19	0.14	0.04		0.06	0.20	0.01	0.01	0.00	0.01	0.01	0.02	0.00	0.00
Xception (FST)	0.00	0.00	0.00	0.00	0.07		0.37	0.18	0.22	0.05	0.29	0.20	0.36	0.11	0.08
ResNet50V2 (FST)	0.00	0.00	0.00	0.00	0.25	0.23		0.11	0.07	0.02	0.12	0.10	0.13	0.03	0.03
InceptionResNetV2 (PFT)	0.00	0.00	0.00	0.00	0.04	0.23	0.41		0.64	0.05	0.83	0.92	0.49	0.20	0.11
ResNet152V2 (PFT)	0.00	0.00	0.00	0.00	0.01	0.42	0.05	0.38		0.06	0.94	0.66	0.35	0.49	0.20
Xception (PFT)	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01		0.29	0.03	0.03	0.24	0.23
ResNet50V2 (PFT)	0.00	0.00	0.00	0.00	0.01	0.04	0.04	0.08	0.20	0.17		0.80	0.62	0.64	0.54
InceptionResNetV2 (CFT)	0.00	0.00	0.00	0.00	0.02	0.40	0.19	0.40	0.30	0.00	0.08		0.49	0.13	0.10
ResNet152V2 (CFT)	0.00	0.00	0.00	0.00	0.02	0.49	0.14	0.76	0.35	0.01	0.08	0.78		0.02	0.01
Xception (CFT)	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.03	0.69	0.11	0.69	0.25	0.00		0.60
ResNet50V2 (CFT)	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.05	0.17	0.09	0.89	0.01	0.01	0.64	

performances, whereas others exhibit greater variability. The performances of some models around 60% in some folds likely reflect model variance, highlighting the sensitivity of high-capacity models to data-specific nuances, which can be mitigated by expanding the training dataset to improve stability, generalization, and robustness of the results. Notably, all PFT and CFT models achieve accuracy rates exceeding 90%, with InceptionResNetV2 models demonstrating particularly consistent performance, surpassing 94% accuracy with only a few outliers.

The boxplot represented in Figure 8 indicates a noteworthy variability is observed in the performance of the statistical and FST models, with AUC scores ranging from approximately 80% to 100%. Outliers are identified across all 15 models analyzed, indicating instances of exceptional performance or potential anomalies. Moreover, the central tendency observed in the AUC scores reinforces the superior performance of transfer learning models compared to their statistical counterparts.

6. CONCLUSIONS AND FUTURE WORK

This study proposes an intelligent fault detection method for identifying fluid leakage in gondola wagon brakes. Leveraging deep transfer learning techniques applied to acoustic signals recorded during wagon brake inspections, this approach demonstrates robustness and reliability in a real-world dataset. The core novelty of this work is the application of deep transfer learning to an industrial pneumatic brake leak detection problem using real inspection data collected under authentic operating conditions. Notably, the method exhibits high accuracy in detecting fluid leakage incidents, enabling timely

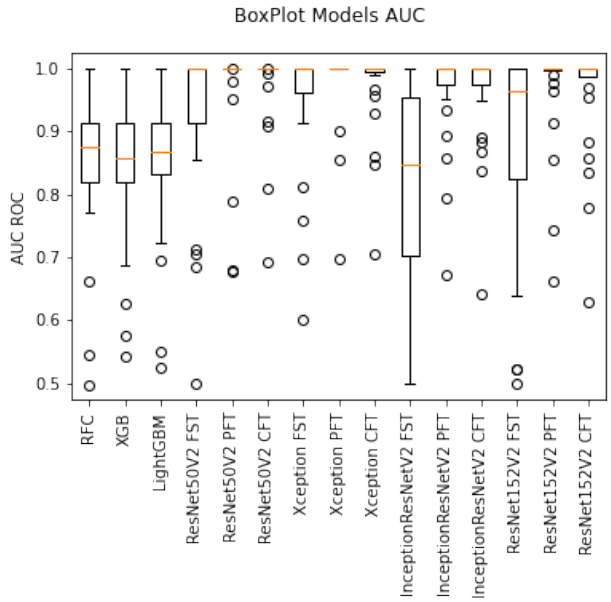

Figure 8. Boxplot of the models AUC performance.

preventive maintenance actions. By harnessing deep transfer learning, it is mitigated the need for large annotated datasets specific to the target domain, overcoming the challenge of data scarcity in fault detection applications. Additionally, utilizing acoustic signals offers a non-intrusive and cost-effective means of monitoring gondola wagon brake health, facilitating continuous monitoring without major operational disruptions. In standard human inspection, trained personnel identify compressed air leaks by listening for characteristic high-frequency hissing sounds, a process that relies on auditory acuity, experience, and selective attention; this reinforces the practical relevance of automating leak detection with the proposed approach.

Through knowledge transfer from pre-trained models and analysis of images generated from audio files, the proposed method outperforms traditional approaches in fault detection. To ground its superiority, it was established a baseline comparison using statistical models on tabular data. Performance evaluation metrics such as accuracy and area under the receiver operating characteristic curve underscore the effectiveness of our method. Furthermore, the validity of results were confirmed through *t-student* and *Wilcoxon* statistical tests.

In conclusion, the presented intelligent fault detection method yields a significant advancement in enhancing the safety and reliability of railway transportation systems. Leveraging state-of-the-art deep transfer learning techniques, it establishes a framework capable of efficiently identifying fluid leakage in gondola wagon brakes based on acoustic signals. The findings demonstrate robust identification of leakage conditions with an accuracy exceeding 94% across the four neural network models with full and partial fine-tuning. This method serves as a valuable auxiliary tool for inspection activities, aiding in the detection of wagon brake fluid leakage through analysis of acoustic signals. Accurate inspections can prevent accidents resulting from insufficient maintenance, thereby saving lives and preserving the public image of companies.

These results, nevertheless, should be interpreted within the scope of the operational setting evaluated in this study. The findings demonstrate robust model performance within the specific operational context studied, with high accuracy and AUC values aligned with the physical characteristics of pneumatic leaks and the adopted mel-spectrogram representation. However, generalization to other scenarios—such as different rail yards, railcar types, or environmental conditions—requires further validation with larger and more heterogeneous datasets. Future studies should therefore include expanded experimental campaigns to assess the robustness and generalization capability of the method across diverse operational contexts.

As a future research direction, integrating this transfer learning method into a robotic system capable of navigating wagon inspection lines could represent a significant advancement. Such a system could streamline and enhance wagon brake inspection processes, potentially reducing or replacing the need for human intervention. Furthermore, since this study dataset was collected in specific conditions as detailed in Section 1, future investigations could explore gathering data under a broader range of weather conditions, including rainy days, or during the operation of adjacent inspection lines. This expanded data collection could be facilitated by utilizing robots, potentially reducing the exposure of humans to safety risks. As the dataset expands, class imbalance may arise. In such cases, the adoption of imbalance handling strategies—such as class weighting, resampling procedures, and synthetic data generation techniques like SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002)—can mitigate training bias, promote more balanced evaluation across classes, and enhance the robustness and reliability of the reported results.

The dataset collected and the code developed as part of this research endeavor are publicly available on GitHub. The dataset includes acoustic signals recorded from gondola wagon brakes, along with corresponding annotations indicating the presence or absence of fluid leakage incidents. This comprehensive dataset serves as a valuable resource for researchers and practitioners interested in exploring fault detection methodologies in railway transportation systems. The code repository contains the implementation of the intelligent fault detection method proposed in this study, facilitating replication and extension of our findings.

ACKNOWLEDGEMENTS

Special thanks are extended to Vale S. A. Wagons Inspection Team for making the inspections accessible for audio recording, and to Christian Bertrand, for his support. Flávio Miguel

Varejão thanks for the financial support of Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq) and Fundação de Amparo à Pesquisa e Inovação de Espírito Santo (Fapes) - TO 1138/2025, 2025-Z9VRS.

DECLARATIONS

- Funding This research did not receive specific grants from funding agencies in the public, commercial or non-profit sectors.
- Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.
- Ethics approval and consent to participate
  The authors declare that this research was conducted with ethics approval and consent to participate.
- Consent for publication The authors consent to this paper publication.
- Data availability
  The data related to this research are available at `https://github.com/jordanaLucia/transferLearning4brakeLeakageDetection`. Access to the data will be provided upon request.
- Code availability
  The code related to this research is available at `https://github.com/jordanaLucia/transferLearning4brakeLeakageDetection`. Access to the code will be provided upon request.
- Author contribution
  First Author: Conceptualization, Methodology, Software, Writing – original draft. **Second Author:** Writing – review & editing. **Last Author:** Supervision, Methodology, Validation, Writing – review & editing.
- Use of AI tools
  The authors acknowledge the use of artificial intelligence to improve the writing and clarity of this manuscript. The authors reviewed the generated text and take full responsibility for the content of the publication.

REFERENCES

Academic Accelerator. (n.d.). *Open wagon - encyclopedia, science news & research reviews.* `https://www.academic-accelerator.com/encyclopedia/open-wagon`. (Accessed: 2024-10-05)

ANTF, Associação Nacional dos Transportadores Ferroviários. (2023). *O setor ferroviário de carga brasileiro.* `https://www.antf.org.br/o-setor-ferroviario-de-carga-brasileiro`. (Accessed: 2024-10-05)

Arunkumar, K. M., & Manjunath, T. C. (2016). A brief review/survey of vibration signal analysis in time domain. *International Journal of Electronics and Communication Engineering*, *3*(3), 12-15. Retrieved from `https://doi.org/10.14445/23488549/IJECE-V3I3P104` doi: 10.14445/23488549/IJECE-V3I3P104

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi: 10.1613/jair.953

Chen, S. (2022). *Deep learning based sound identification.* `http://ens.ewi.tudelft.nl/islandora`.

Chen, Y., Guo, Q., Liang, X., Wang, J., & Qian, Y. (2019). Environmental sound classification with dilated convolutions. *Applied Acoustics*, *148*, 123–132.

Chen, Z., Xia, J., Li, J., Chen, J., Huang, R., Jin, G., & Li, W. (2023). Generalized open-set domain adaptation in mechanical fault diagnosis using multiple metric weighting learning network. *Advanced Engineering Informatics*, *56*, 101781. doi: 10.1016/j.aei.2023.101781

Cheng, C., Qiao, X., Luo, H., Teng, W., Gao, M., Zhang, B., & Yin, X. (2019). A semi-quantitative information based fault diagnosis method for the running gears system of high-speed trains. *IEEE Access*, *7*, 38168-38178. Retrieved from `https://doi.org/10.1109/ACCESS.2019.2901022` doi: 10.1109/ACCESS.2019.2901022

Deng, J., Dong, W., Socher, R., Li, L., Kai, L., & Li, F. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Ge, X., Chen, Q., Ling, L., Zhai, W., & Wang, K. (2023). An approach for simulating the air brake system of long freight trains based on fluid dynamics. *Railway Engineering Science*, *31*(2), 122–134. Retrieved from `https://link.springer.com/article/10.1007/s40534-022-00291-0` doi: 10.1007/s40534-022-00291-0

Giorgi, B. D., Levy, M., & Apple, R. S. (2022). Mel spectrogram inversion with stable pitch. *Apple Machine Learning Research*. Retrieved from `https://machinelearning.apple.com/research/mel-spectrogram`

Godbole, V., Dahl, G. E., Gilmer, J., Shallue, C. J., & Nado, Z. (2022). *Deep learning tuning playbook.* `https://github.com/google-research/tuning_playbook`.

Hashmi, M., Ibrahim, M., Bajwa, I., Siddiqui, H.-U.-R., Rustam, F., Lee, E., & Ashraf, I. (2022). Railway track inspection using deep learning based on audio to spectrogram conversion: An on-the-fly approach. *Sensors*, *22*(5), 1983. Retrieved from `https://`

www.mdpi.com/1424-8220/22/5/1983 doi: 10.3390/s22051983

Hoang, N. T., & Kim, D. (2022). Supervised contrastive resnet and transfer learning for the in-vehicle intrusion detection system. *Expert Systems with Applications*, *209*, 118390. doi: 10.1016/j.eswa.2022.118390

Hu, R., Zhang, M., Meng, X., & Kang, Z. (2022). Deep subdomain generalisation network for health monitoring of high-speed train brake pads. *Engineering Applications of Artificial Intelligence*, *113*, 104896. doi: 10.1016/j.engappai.2022.104896

Jin, Y., Xie, G., Li, Y., Zhang, X., Han, N., Shangguan, A., & Chen, W. (2021). Fault diagnosis of brake train based on multi-sensor data fusion. *Sensors*, *21*(13), 4370. Retrieved from https://doi.org/10.3390/s21134370 doi: 10.3390/s21134370

Karabacak, Y. E., Özmen, N. G., & Gümüşel, L. (2022). Intelligent worm gearbox fault diagnosis under various working conditions using vibration, sound and thermal features. *Applied Acoustics*, *186*, 108463. Retrieved from https://www.sciencedirect.com/science/article/pii/S0003682X21005570 doi: 10.1016/j.apacoust.2021.108463

Kensert, A., Harrison, P. J., & Spjuth, O. (2018, October). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. In *Uppsala university faculty of pharmacy 50th anniversary.* Uppsala University. Retrieved from https://pharmb.io/poster/2018-transfer-cnn/ doi: 10.6084/m9.figshare.9747029.v1

Kulkarni, R., Qazizadeh, A., & Berg, M. (2023). Unsupervised rail vehicle running instability detection algorithm for passenger trains (ivrida). *Measurement*, *217*, 112372. Retrieved from https://www.sciencedirect.com/science/article/pii/S026322412300122X doi: 10.1016/j.measurement.2023.112372

Liu, H., Peng, J., Gao, D., Yang, Y., Fan, Y., Hu, C., & Zhang, X. (2020). A hybrid data-fusion estimate method for health status of train braking system. In *Proceedings of the 2020 ieee international conference on systems, man, and cybernetics (smc)* (pp. 2130–2135). Toronto, ON, Canada: IEEE. doi: 10.1109/SMC42913.2020.9281152

Luz, J. S., Oliveira, M. C., Araújo, F. H. D., & Magalhães, D. M. V. (2021). Ensemble of handcrafted and deep features for urban sound classification. *Applied Acoustics*, *175*, 107819. Retrieved from https://www.sciencedirect.com/science/article/pii/S0003682X20309245 doi: 10.1016/j.apacoust.2020.107819

Ma, S., Leng, J., Chen, Z., Li, B., Li, X., Zhang, D., ... Liu, Q. (2024). A novel weakly supervised adversarial network for thermal error modeling of electric spindles with scarce samples. *Expert Systems with Applications*, *238*, 122065. doi: 10.1016/j.eswa.2023.122065

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *In proceedings of the 14th python in science conference* (pp. 18–25). SciPy.

Padha, A., & Sahoo, A. (2023). Qclr: Quantum-lstm contrastive learning framework for continuous mental health monitoring. *Expert Systems with Applications*, *226*, 120761. doi: 10.1016/j.eswa.2023.120761

Patil, S. S., Pardeshi, S. S., & Patange, A. D. (2023). Health monitoring of milling tool inserts using cnn architectures trained by vibration spectrograms. *CMES - Computer Modeling in Engineering and Sciences*, *139*(2), 487-500. Retrieved from https://www.techscience.com/cmes/v139n2/49963 doi: 10.32604/cmes.2023.022924

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pfaff, R., Enning, M., & Sutter, S. (2024). A risk-based approach to automatic brake tests for rail freight service: incident analysis and realisation concept. *Journal of Rail Transport*, *15*(2), 123-145.

Pugi, L., Rosano, G., Viviani, R., Cabrucci, L., & Bocciolini, L. (2023). Modeling, testing and validation of the vibrational behavior of a dynamometric test rig for railway braking systems. *World Journal of Engineering*, 1–18. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/WJE-04-2023-0073/full/html doi: 10.1108/WJE-04-2023-0073

Rauber, T., Loca, A. L. S., Boldt, F. A., Rodrigues, A. L., & Varejão, F. M. (2021). An experimental methodology to evaluate machine learning methods for fault diagnosis based on vibration signals. *Experts Systems with Applications*, *167*, 114022. doi: https://doi.org/10.1016/j.eswa.2020.114022

Reis, J. L., & Varejao, F. M. (2023). Detecç ao de vazamentos de fluidos de freios a ar em vagões do tipo gôndola através do sinal acústico: Um modelo de classificação de falhas. *Revista Foco*, *16*(5), e1885. Retrieved from https://doi.org/10.54751/revistafoco.v16n5-072 doi: 10.54751/revistafoco.v16n5-072

Ribeiro, R. P., Pereira, P., & Gama, J. (2016). Sequential anomalies: a study in the railway industry. *Machine Learning*, *105*(1), 127–153. doi: 10.1007/s10994-016-5584-0

Rodrigues, P. L. C., Jutten, C., & Congedo, M. (2023). Riemannian procrustes analysis: Transfer learning for brain-

computer interfaces. *IEEE Transactions on Biomedical Engineering*, *70*(9), 2715-2725. doi: 10.1109/TBME.2023.3235513

Sakellariou, J. S., Petsounis, K. A., & Fassois, S. D. (2014). Vibration based fault diagnosis for railway vehicle suspensions via a functional model based method: A feasibility study. *Journal of Mechanical Science and Technology*, *28*(11), 4443–4451. doi: 10.1007/s12206-014-1019-2

Sangeetha, P., & Hemamalini, S. (2017). Dyadic wavelet transform-based acoustic signal analysis for torque prediction of a three-phase induction motor. *IET Signal Processing*, *11*(5), 604–612. doi: 10.1049/iet-spr.2016.0165

Tagawa, Y., Maskeliūnas, R., & Damaševičius, R. (2021). Acoustic anomaly detection of mechanical failures in noisy real-life factory environments. *Electronics (Switzerland)*, *10*(19), 2329. doi: 10.3390/electronics10192329

Umesh, S., Cohen, L., & Nelson, D. (1999). Fitting the mel scale. In *Icassp, ieee international conference on acoustics, speech and signal processing - proceedings* (pp. 217–220). IEEE.

Ustubioglu, A., Ustubioglu, B., & Ulutas, G. (2023). Mel spectrogram-based audio forgery detection using cnn. *Signal, Image and Video Processing*, *17*(5), 1915-1924. doi: 10.1007/s11760-022-02274-y

Vrbančič, G., & Podgorelec, V. (2020). Transfer learning with adaptive fine-tuning. *IEEE Access*, *8*, 196197–196211. Retrieved from https://doi.org/10.1109/ACCESS.2020.3034343 doi: 10.1109/ACCESS.2020.3034343

Wang, X., Lu, Z., Wei, J., & Zhang, Y. (2019). Fault diagnosis for rail vehicle axle-box bearings based on energy feature reconstruction and composite multiscale permutation entropy. *Entropy*, *21*(9), 865. Retrieved from https://doi.org/10.3390/e21090865 doi: 10.3390/e21090865

Wood, L., Tan, Z., Stenbit, I., Bischof, J., Zhu, S., & Chollet, F. (2022). *Kerascv*. https://github.com/keras-team/keras-cv.

Wu, J. D., & Chan, J. J. (2009). Faulted gear identification of a rotating machinery based on wavelet transform and artificial neural network. *Expert Systems with Applications*, *36*(5), 8862–8875. doi: 10.1016/j.eswa.2008.11.035

Xiong, L., Liu, Z., & Niu, G. (2021). Failure analysis of electromagnetic-pneumatic valve in bogie-based braking control system. In *2021 global reliability and prognostics and health management, phm-nanjing 2021*.

Xu, M., & Yao, H. (2023). Fault diagnosis method of wheelset based on eemd-mpe and support vector machine optimized by quantum-behaved particle swarm algorithm. *Measurement*, *216*, 112923. doi: 10.1016/j.measurement.2023.112923

Yao, Y., Wang, H., Li, S., Liu, Z., Gui, G., Dan, Y., & Hu, J. (2018). End-to-end convolutional neural network model for gear fault diagnosis based on sound signals. *Applied Sciences (Switzerland)*, *8*(9), 1584. doi: 10.3390/app8091584

Ye, Y., Zhang, J., & Liang, H. (2020). An acoustic-based recognition algorithm for the unreleased braking of railway wagons in marshalling yards. *IEEE Access*, *8*, 78038–78048. doi: 10.1109/ACCESS.2020.3006003

Yusof, M. F. M., Kamaruzaman, M. A., Ishak, M., & Ghazali, M. F. (2017). Porosity detection by analyzing arc sound signal acquired during the welding process of gas pipeline steel. *International Journal of Advanced Manufacturing Technology*, *89*(9-12), 3661-3670. doi: 10.1007/s00170-016-9343-4

Zanelli, F., Mauri, M., Castelli-Dezza, F., Sabbioni, E., Tarsitano, D., & Debattisti, N. (2022). Energy autonomous wireless sensor nodes for freight train braking systems monitoring. *Sensors*, *22*(5), 1876. Retrieved from https://doi.org/10.3390/s22051876 doi: 10.3390/s22051876

Zhang, M., Liu, Z., & Dang, X. (2021). Fault diagnosis on train brake system based on multi-dimensional feature fusion and gbdt enhanced classification. *Sensors*, *21*(13), 4370. doi: 10.3390/s21134370

Zhang, T., Peng, P., Tang, X., Yan, R., & Deng, R. (2023). Cme-epc: A coarse-mechanism embedded error prediction and compensation framework for robot multi-condition tasks. *Robotics and Computer-Integrated Manufacturing*, *78*, 102436. doi: 10.1016/j.rcim.2022.102436

Zhou, D., Ji, H., He, X., & Shang, J. (2018). Fault detection and isolation of the brake cylinder system for electric multiple units. *IEEE Transactions on Control Systems Technology*, *26*(5), 1744–1757. doi: 10.1109/TCST.2018.2794716

Zhou, Q., Shan, J., Ding, W., Wang, C., Yuan, S., Sun, F., ... Fang, B. (2021). Cough recognition based on mel-spectrogram and convolutional neural network. *Frontiers in Robotics and AI*, *8*, 580080. doi: 10.3389/frobt.2021.580080

Zuo, J., Ding, J., & Feng, F. (2019). Latent leakage fault identification and diagnosis based on multi-source information fusion method for key pneumatic units in chinese standard electric multiple units (emu) braking system. *Applied Sciences*, *9*(2), 300. doi: 10.3390/app9020300