Using Vehicle Interior Noise Classification for Monitoring Urban Rail Transit Infrastructure

Yifeng Wang [1] , Ping Wang [1], Qihang Wang [1], Zhengxing Chen [1] and Qing He [1,2,3,*]

[1] Key Laboratory of High-Speed Railway Engineering of the Ministry of Education, School of Civil Engineering, Southwest Jiaotong University, Chengdu 610031, China; yfw@my.swjtu.edu.cn (Y.W.); wping@swjtu.edu.cn (P.W.); qihangwang@my.swjtu.edu.cn (Q.W.); chenzhengxing@my.swjtu.edu.cn (Z.C.)

[2] Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA

[3] Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA

* Correspondence: qinghe@buffalo.edu

check for updates

Abstract: This study developed a multi-classification model for vehicle interior noise from the subway system, collected on smartphones. The proposed model has the potential to be used to analyze the causes of abnormal noise using statistical methods and evaluate the effect of rail maintenance work. To this end, first, we developed a multi-source data (audio, acceleration, and angle rate) collection framework via smartphone built-in sensors. Then, considering the Shannon entropy, a 1-second window was selected to segment the time-series signals. This study extracted 45 features from the time- and frequency-domains to establish the classifier. Next, we investigated the effects of balancing the training dataset with the Synthetic Minority Oversampling Technique (SMOTE). By comparing and analyzing the classification results of importance-based and mutual information-based feature selection methods, the study employed a feature set consisting of the top 10 features by importance score. Comparisons with other classifiers indicated that the proposed XGBoost-based classifier runs fast while maintaining good accuracy. Finally, case studies were provided to extend the applications of this classifier to the analysis of abnormal vehicle interior noise events and evaluate the effects of rail grinding.

Keywords: urban rail transit interior noise; smartphone sensing; XGBoost classifier; railway maintenance

1. Introduction

By the end of 2018, the total operating mileage of urban rail transit (URT) in China exceeded 5700 km, including 4350 km of subway lines, and it is expected to double in the next 3 to 5 years [1]. With the rapid extension of the URT network, the current maintenance mode relies on humans, and it is challenging to ensure the safe and stable operation of trains. Therefore, intelligent URT maintenance work should be promoted for higher efficiency.

As one of the most prevalent kinds of URT, subways are increasingly essential in people's daily lives. However, abnormal vibration and noise significantly affect passengers' riding experience. Moreover, these abnormalities provide information about wheel-rail interactions and degradation of the track structures. Generally, train-induced noise can be categorized as external or interior noises [2]. Vehicle interior noise which is pertinent to this study mainly consists of noise from electrical equipment, aerodynamic noise, and wheel-rail noises [3]. Usually, the aerodynamic noise is dominant when the train speed exceeds 250 km/h, and electrical equipment noise dominates for speeds slower than

35 km/h [4]. As the subway trains usually run at 30–80 km/h, the wheel-rail noise is the main component of vehicle interior noise [5]. The wheel-rail interaction significantly influences the wheel-rail noise. Therefore, we assumed that there exists a mapping relationship between vehicle interior noises and wheel-rail interactions. This mapping relationship provides an approach to monitor track conditions through vehicle interior noise. Moreover, it would be convenient to develop a simple onboard interior noise monitoring system that contributes to the safety and reliability of the railway system.

Regarding vehicle interior noise, past studies have mainly focused on the generation mechanism, transmission characteristics, and control strategies [6–10]. Typical study topics, such as noise characteristics analysis [11], sound quality evaluation [12], and noise level prediction [13], can be attributed to the above research fields. However, because the vehicle-track coupling system consists of a large number of components, the interior noise is affected by numerous factors, such as track slab [14], rail roughness, wheel out-of-roundness [9], and car body structure [15]. These factors may interact with each other and influence the characteristics of vehicle interior noise. Therefore, researchers generally choose one or two factors, such as rail fastener stiffness [7] and wheel polygonal wear [9], to perform their analysis at a lower complexity.

Among related studies, the prediction of vehicle interior noise is one of the most prevalent topics because it benefits the design and construction of track-vehicle systems at the early stages. Methods such as the boundary element method (BEM) [16], finite element method (FEM) [17], and statistical energy analysis method (SEAM) [15] are commonly used in this. However, their effectiveness relies significantly on the selected boundary conditions and model parameters. Thus, these numerical models are generally applied for specific problems. Moreover, the results of field tests are also often used for model verification. Despite the effectiveness of the method combining analytical models, numerical simulation, and field tests in the study of vehicle interior noise, the difficulty to obtain model parameters limits its application. Moreover, field tests may also interfere with daily operations. Overall, these studies do not make the best use of data collected during the daily operation and maintenance of the railway system.

In this context, the railway transportation industry is at the forefront of implementing analytics and big data [18]. Machine learning (ML) and artificial intelligence (AI) are two concepts at the leading edge of information technology, both of which contribute to big data technology. In recent years, the implementation of ML in the railway industry has been widely studied, for example in the prediction of passenger flow [19], delay events [20], and railway operation disruptions [10]. Moreover, many cases have been reported for railway infrastructure management and maintenance, including the detection and diagnosis of defects [21–23], prediction of failure events [24,25], and forecast of remaining useful life of devices [26]. These studies indicate that ML technologies have a promising prospect in promoting intelligent railway maintenance, thus ensuring the safety of the railway transit system.

As for data on vehicle interior noise, users require automatic methods to segment, label, and store the increasing amount of acoustic data from monitoring systems. The major challenge in this field is the automatic classification of audio [27]. Recent studies on the classification of traffic noise have been conducted, for example, to identify the type of vehicle through roadside noise [28,29] and evaluate passengers' subjective experience by categorizing the cabin's interior noise [30]. However, compared with traffic noise, the factors influencing vehicle interior noise of subway trains are considerably more complicated.

For collecting track conditions, the railway industry has employed various dedicated devices, such as track inspection vehicles [31] and visual inspection systems [32]. Although these devices perform well in detecting track conditions, the expensive cost and the interference for regular operation limit their usage in urban rail transit systems. There are also some on-board devices being developed to monitor track conditions using in-service vehicles [33–35]. However, the installation of these devices may change the design characteristics of cars and cause potential safety issues. As of now, these novel on-board monitoring devices have not been widely used. As an integrated platform, a smartphone can achieve data collection, storage, and transmission individually. Besides, the smartphone is mature,

cost-effective, and easy to use, promoting its application in various fields. Studies using the embedded accelerometers of smartphones to monitor road conditions and evaluating the ride quality have been reported [36,37]. These research works inspired the authors to investigate the feasibility of using smartphones to collect multi-source data about subway vehicles.

According to the above literature review, current studies about vehicle interior noise mainly focus on its generation mechanism and influencing factors through analytical models, numerical simulations, and field tests. To the best of our knowledge, only a few studies have analyzed vehicle interior noise using data-driven methods. Therefore, this study aims to advance data mining of vehicle interior noise for decision making in rail maintenance, such as for rail grinding. In this context, there are two significant challenges. First, despite sensing technologies being well developed now, it is still difficult to establish an onboard data collection framework that is easy to deploy, cost-efficient, and reliable. Moreover, the simultaneous collection of dynamic responses from the car body and interior noise is essential because these two datasets are connected to each other. Second, due to the complexity of vehicle interior noise, the extraction of useful features and correct labeling of noise classes remain challenging.

The goal of this study is to mine useful information from the vast amount of interior noise data using ML methods. To pursue this goal, onboard smartphone data were collected, including dynamic responses and noises. Further, a series of analyses were performed to classify the noises and clarify the influencing factors. The novel contributions of this paper are summarized as follows:

1. A smartphone-based onboard data collection framework for vehicle interior noise and dynamic responses of the car body was established.
2. The theory of Shannon entropy was considered when selecting the optimal window size for segmenting the multi-source time-series signals.
3. A multi-classification model for subway vehicle interior noise was established based on the XGBoost algorithm. The generation of a set of 45 features and performing feature selection based on different methods were also included.
4. Case studies were conducted to extend the application scenario for the analysis of abnormal noise causes and evaluating the effect of rail grinding.

This paper is organized as follows. Section 2 briefly illustrates the research methodology. Section 3 introduces the data utilized in this study and its collection framework. Section 4 describes the modeling approaches, including data segmentation and time windows, and establishes the multi-classification model with the Extreme Gradient Boosting (XGBoost) method. Furthermore, Section 5 presents the analysis results and discussions. Finally, in Section 6, conclusions are drawn according to the relevant analysis.

2. Research Methodology

The research methodology of this study is shown in Figure 1. First, we developed an Android app that leverages built-in sensors of onboard smartphones to collect vehicle interior noise and the corresponding dynamic responses of the car body. Second, time windows were used to segment the multi-source signals and establish the corresponding relationship between the audio and other signals. This method was significantly effective in overcoming the difficulty brought by the different sampling frequencies of a variety of sensors. Third, features were generated and selected from the time- and frequency-domains. Fourth, an automatic classification model for train interior noise was developed using XGBoost, a tree-based method. Finally, the proposed model was validated based on field experiments on the subway line.
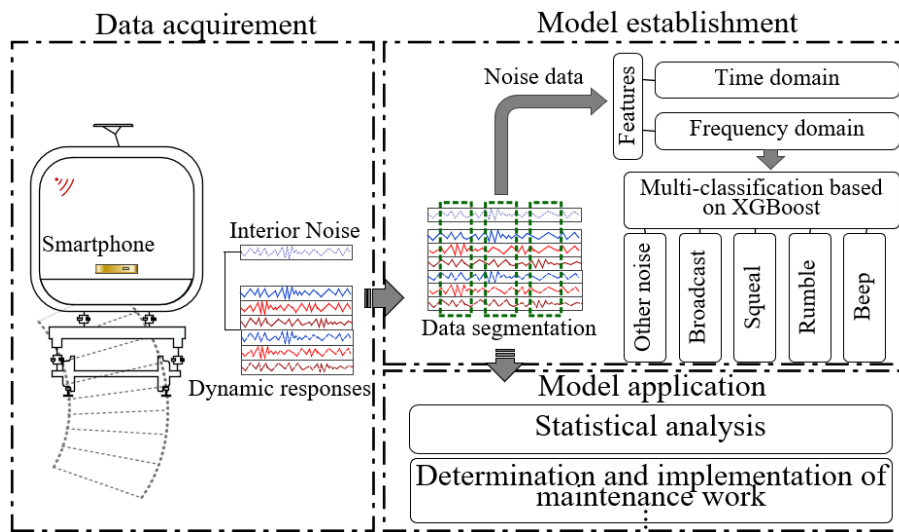
Figure 1. Research methodology of this study.

3. Data Collection and Description

Figure 2 shows the field test setup for data collection using Android smartphones (Huawei Honor FRD-AL00). During the test, the smartphone was placed on the cabin floor, right above the bogie to sense the response from the wheel-rail contact interface. In a parallel study, we verified that the differences between smartphone sensors and high-precision industry accelerators are acceptable, especially in the vertical direction [36]. Thus, the dynamic response signals can be considered a good record of the movement state of the car body. An app was developed to save and transmit the data to our cloud server. In the field test, three sensors were used, namely the microphone, accelerometer, and gyroscope. Moreover, considering the performance of these sensors and the characteristics of the signals, the sampling frequency of the accelerometer and gyroscope were set to 100 Hz, and that of the microphone to 22,050 Hz.
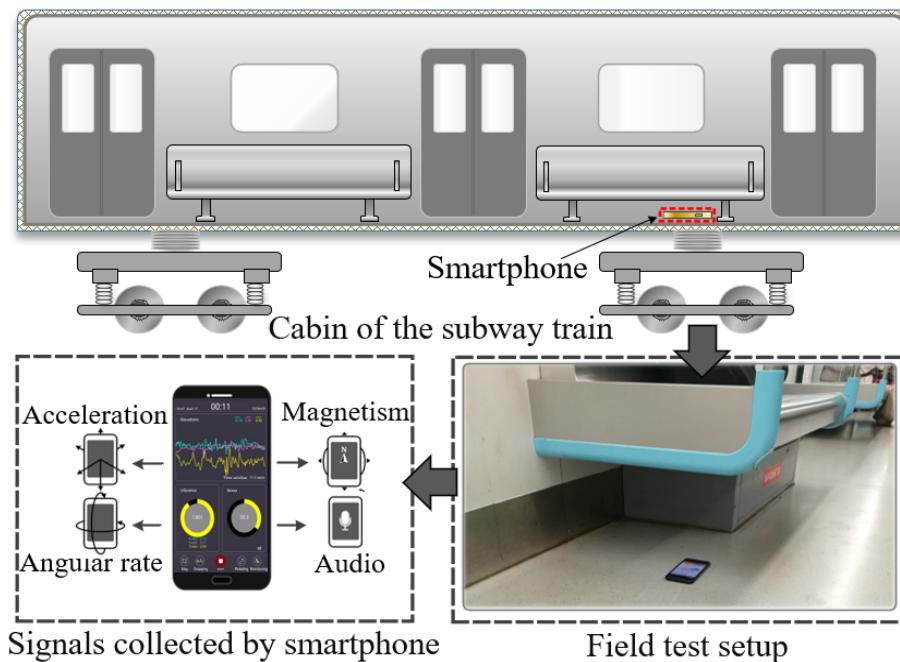

Figure 2. Data collection with the smartphone.

In this study, all tests were performed on Line 7 of the Chengdu Metro, China, which is a loop subway line. Its layout is shown in Figure 3a. This line covers 38.61 km and 31 stations, and it started operations in December 2017. The trains run along the outer and inner loop, with a maximum speed of 80 km/h. Because this is a loop line, it contains a large number of curve sections (166 curves). The radius distribution of these curves is presented in Figure 3b. It is challenging to maintain the track structures in good conditions due to the high number of curves, and the squeal that typically occurs along the curves is one of the most significant problems.
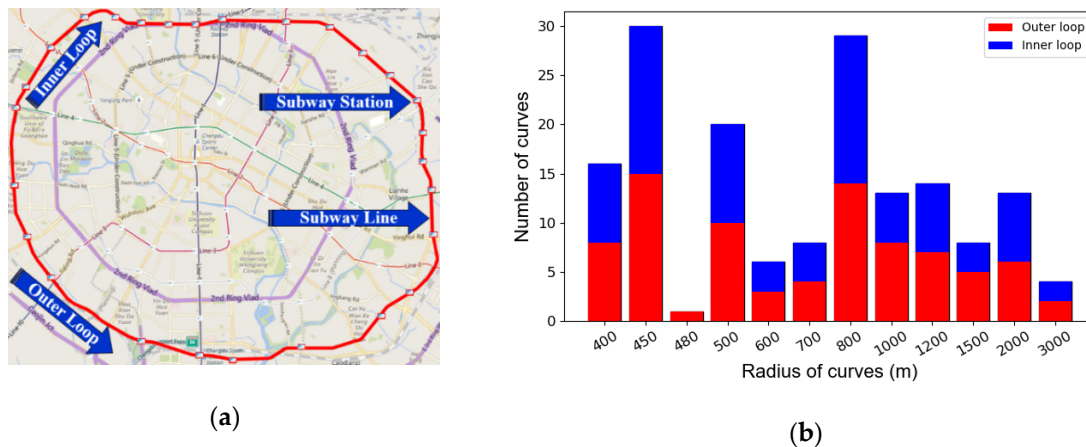

(**a**)   (**b**)

Figure 3. Line 7 of the Chengdu Metro, China: (**a**) Overview; (**b**) Radius of curves.

The data used in this study were collected on 2 August 2019, and 1 October 2019, before and after rail grinding. There were more abnormal events in the dataset before rail grinding. The data from August was used to train and test the multi-classification model, and to justify the need for rail grinding. The data measured on both days were compared. When training the model, we manually labeled the audio sequence into five groups, including 'Other noises', 'Broadcast', 'Squeal', 'Rumble', and 'Beep'. Here, 'Broadcast' refers to the official broadcast by the subway system or passengers' voices. 'Squeal' is an intense noise generated by the relative movement between wheel and rail. 'Rumble' refers to a low heavy sound when the train passes a specific area. 'Beep' is the alarm sound when a door is opened or closed. 'Other noises' refers to a sound which cannot be categorized into the above four classes. The time-frequency characteristics of these five classes of noise are presented in Figure 4.
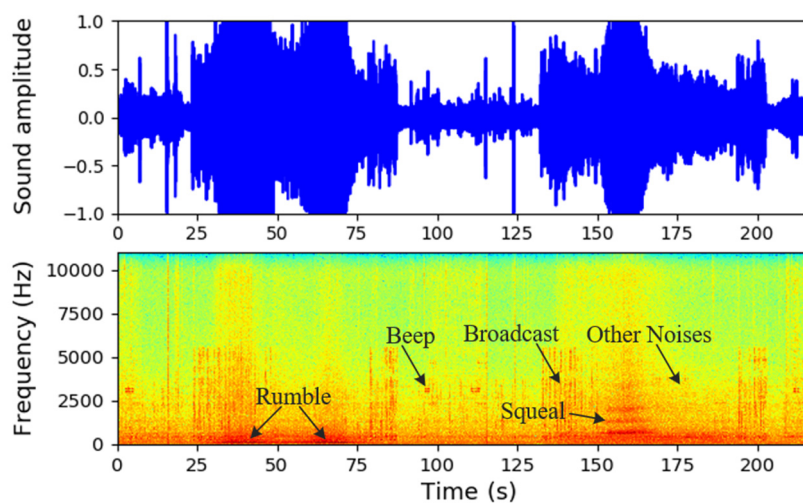

Figure 4. Data collection with the smartphone.

4. Model Approach

4.1. Data Segmentation and Time Window

Differences in sensor sampling frequencies make it difficult to identify the corresponding relationship among the multi-source signals. In this context, data segmentation is a typical method to preprocess continuous data and capture embedded features. This approach has been frequently implemented in activity recognition, such as in speech [38] and human activity [39] recognition. Therefore, we adopted the moving time-window method to segment the signals in our study. During data segmentation, there were two crucial parameters to be determined the size of the time window and the overlap between two adjacent windows. To avoid the duplication of data interference with statistical analysis, the overlap parameter was set to 0. That is, there was no overlap between two adjacent windows. Although the window method is normally used in data segmentation, there is no clear consensus on which window size should be employed [39]. The characteristics of vehicle interior noise are different from other audio signals. Therefore, we cannot use the window sizes used in speech recognition as a reference. Generally, small windows allow for on-point activity detection with a few resources and low energy costs. In contrast, large windows are usually considered to identify complex activities. To obtain the optimal window size for vehicle interior noise multi-classification, we leveraged the Shannon entropy and the actual requirements when labeling the training data manually.

We assumed that under the optimal window size, the system carries more information than under other situations [40]. The Shannon entropy is a method commonly used to describe the average information of a system, and it can be written as:

$$H = -\sum_{i=1}^{m} p(x_i) \log_2 p(x_i),$$ (1)

where x_i denotes the ith event; m represents the total number of events; and $p(x_i)$ is the probability when $x = x_i$ and $\sum_{i=1}^{m} p(x_i) = 1$. To obtain the optimal window size, the vehicle interior noise signal was first divided into a series of segment sequences according to different window sizes. The standard deviation of each segment was calculated to describe the state of the segment. Consequently, standard deviation sequences corresponding to different window sizes were available. It was then assumed that all values of standard deviation fall within the range of $(0, A]$, where A is the maximum standard deviation under different window sizes. After that, this interval was equally divided into m sub-intervals, where the ith sub-interval can be written as $(a_i, a_{i+1}]$, $a_1 = 0$, and $a_{m+1} = A$. Thus, the optimization model for time window size can be described as:

$$\max H(n) = -\sum_{i=1}^{m} p_i(n) \log_2 p_i(n),$$ (2)

where n is the time window size, and $p_i(n)$ is the probability of standard deviation values to fall into the range of $(a_i, a_{i+1}]$ when the time window size is n . In this study, the optimal time window size was obtained from an extensive number of samples. The size of the windows ranged from 0.1 to 64 s, and the total number of samples was 200. For a higher classification accuracy, more attention should be paid to small windows. To obtain those samples, logarithm interpolation was used. For all samples, the next sample is always $10^{(\log_{10} 64 - \log_{10} 0.1)/200}$ times the previous one. By calculating the Shannon entropy considering all 200 sizes, we obtained the maximum entropy and its corresponding window size.

4.2. Data Balance Using the Synthetic Minority Oversampling Technique (SMOTE)

The pie chart in Figure 5a shows the proportion of the five categories of vehicle interior noise studied in this work. The most frequent event is 'Broadcast', which accounts for 67.56% of all vehicle

interior noise events. 'Other noises' is the next most frequent event, at approximately 22%. 'Beep', 'Squeal', and 'Rumble' represent smaller percentages of the vehicle interior noise events, at 4.99%, 2.79%, and 2.66%, respectively. These results indicate that there is a severe class imbalance, which could significantly undermine most standard classification learning algorithms [41].
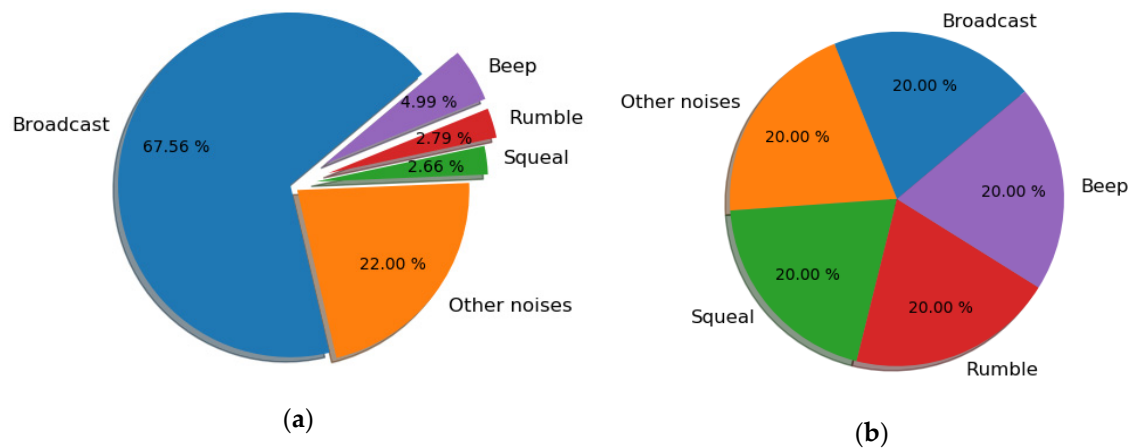

Figure 5. Data (**a**) before and (**b**) after synthetic minority oversampling technique (SMOTE) balance.

In this study, we adopted the synthetic minority oversampling technique (SMOTE) to overcome data imbalance. Generally, the class imbalance can be addressed by: (1) synthesizing new minority class instances; (2) oversampling minority class; (3) under-sampling majority class; and (4) tweaking the cost function to enhance the importance of misclassification of minority instances. The SMOTE used in this study utilizes the first solution because increasing the number of minority classes is better than merely duplicating minority classes, which has stronger robustness and generalization ability. This technique returns the original samples and an additional number of synthetic minority class samples. The SMOTE takes samples from the feature space of each minority class and its k nearest neighbors and generates new instances that combine the features of the target classes with the features of their k neighbors. Therefore, it increases the features available for each category and makes the samples more general. In this study, we increased the percentage of 'Other noises', Squeal', 'Rumbel', and 'Beep' to be the same as 'Broadcast' via SMOTE when training the multi-classification model, as shown in Figure 5b.

4.3. Features

In ML, features are individual measurable properties of an observed phenomenon [42]. Selecting informative, independent, and discriminating features is a crucial process in classification or regression. The 45 features implied in this study are shown in Table 1. The feature sets include low-level signal properties (f1–f9) and Mel-frequency spectral coefficients (MFCCs) (f10–f45) [27].

Table 1 defines the features of low-level signal properties (f1–f9). N is the sample number of one segment; k refers to the kth sample point; x is the time-series signal; and X denotes the spectrum of Fourier transform (FT); $sign(\ )$ is the sign function; TH is the threshold, which takes the value of 0.85 in the definition of f6; $P(k)$, which is shown in the definition of f8, is the probability distribution of the power spectrum $S(k) = |X(k)|^2$. Moreover, MFCCs are features commonly used in speech and speaker recognition [38]. In this study, the first 12 MFCCs coefficients (f10–f21) were used to obtain more information from the audio segments. Because the audio signals vary intermittently, it is necessary to add features related to the change of cepstral characteristics over time [43]. Therefore, the first- and second-order derivatives of the first 12 MFCCs (f22–f33 and f34–f45) were also calculated.

<div align="center">

Table 1. Features used in this study.

</div>

Category	Feature		Definition
Time-domain	f1	Segment energy	$f1 = \sum\limits_{k=0}^{N-1} \left|x(k)\right|^2$
	f2	Root mean square (RMS) of the segment	$f2 = \sqrt{\frac{1}{N} \sum\limits_{k=0}^{N-1} x(k)^2}$
	f3	Zero cross rate	$f3 = \frac{1}{2} \sum\limits_{k=0}^{N-1} \left|sign(x(k)) - sign(x(k-1))\right|$
Frequency-domain	f4	Spectral centroid	$f4 = \sum\limits_{k=0}^{N-1} \left|X(k)\right| \cdot k / \sum\limits_{k=0}^{N-1} \left|X(k)\right|$
	f5	Spectral bandwidth	$f5 = \sqrt{\sum\limits_{k=0}^{N-1} (k - f4)^2}$ [29]
	f6	Spectral roll-off	$f6 = max\left\{ \sum\limits_{k=0}^{m} \left|X(k)\right| \leq TH \cdot \sum\limits_{k=0}^{N-1} \left|X(k)\right| \right\}$
	f7	Spectral bandwidth to energy ratio	$f7 = f5/f1$
	f8	Spectral entropy	$f8 = - \sum\limits_{n=1}^{N} P(k) \log_2 P(k)$
	f9	Energy to spectral entropy ratio	$f9 = f10/f8$
	f10–f21	First 12 MFCCs	
	f22–f33	First-order derivatives of f10–f21	
	f34–f45	Second-order derivatives of f10–f21	

4.4. Feature Selection Based on IG

During data analysis, hundreds of features may be generated, many of which are redundant and not relevant to the data mining task. Removing these irrelevant features may waste vast amounts of computation time and influence the prediction results. Although experts in relevant files can select the useful features, this is a challenging and time-consuming task, especially when the characteristics of the dataset are not well known. The goal of feature selection is to find a minimum set of features so that the prediction results are as close as possible to (or better than) the original feature set.

In this study, we employed the IG as an index for feature selection. IG is a feature evaluation method based on entropy and is widely employed in the field of ML [44]. In feature selection, IG is defined as the complete information provided by the features for the classification task. IG measures the importance of features as:

$$IG(S, a) = E(S) - E(S|a), \tag{3}$$

where $IG(S, a)$ is the IG of the original feature set S for feature a; $E(S)$ is the entropy for the feature set without any change; and $E(S|a)$ is the conditional entropy for the feature set, given feature a. The conditional entropy $E(S|a)$ can be written as:

$$E(S|a) = \sum\limits_{v \in a} \frac{Sa(v)}{S} * E(Sa(v)), \tag{4}$$

where $\frac{Sa(v)}{S}$ is the categorical probability distribution of feature a at $v \in a$, and $E(Sa(v))$ is the entropy of a sample group where a has the value v. The greater the value of $IG(S, a)$, the more critical is a for the classification model.

4.5. Multi-Classification Model for Vehicle Interior Noise Based on XGBoost

XGBoost was designed based on gradient boosted decision trees [45]. We chose XGBoost due to its computation speed and model performance, which have been verified by a previous study [22]. As an ensemble model of decision trees, the definition of the XGBoost model can be written as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \tag{5}$$

where K is the total number of decision trees, f_k is the kth decision tree, and $\hat{y}_i$ is the prediction result of sample x_i. The cost function with a regularization term is given by [45]:

$$L(f) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{6}$$

with:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2, \tag{7}$$

where T is the number of leaves of the classification tree f, and w is the score of each leaf. The Lasso regulation of coefficient γ and ridge regularization of coefficient λ can work together to control the complexity of the model. By expressing the objective function as a second-order Taylor expansion, the objective function at step t can be written as [46]:

$$L(f) \approx \sum_{i=1}^{n} \left[l(\hat{y}_i, y_i) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{8}$$

where $g_i = \partial_{\hat{y}} l(\hat{y}_i, y_i)$, and $g_i = \partial_{\hat{y}}^2 l(\hat{y}_i, y_i)$. By removing the constant term, the approximation of the objective at step t is available:

$$\hat{L}(f) = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t). \tag{9}$$

By expanding the regularization term Ω and defining I_j as the instance set at leaf j, Equation (9) can be rewritten as [47]:

$$\hat{L}(f) = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T. \tag{10}$$

By rewriting the objective function as a unary quadratic function of leaf score w, the optimal w and the value of the objective function are easily obtained. In XGBoost, the gain is used for splitting decision trees:

$$G_j = \sum_{i \in I_j} g_i, \tag{11}$$

$$H_j = \sum_{i \in I_j} h_i, \tag{12}$$

$$gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma, \tag{13}$$

where the first and second terms are the score of the left and right child tree, respectively; the third term is the score if there is no splitting; and γ is the complexity cost when a new split is added. Despite

the serial relationship between the adjacent trees, the node in a certain level can be parallel during the splitting, which enables XGBoost to have a faster train speed.

5. Results and Discussions

In general, the parameters of an ML model can significantly impact its performance, and XGBoost is no exception. Through extensive testing and observation, we set the critical parameters of this model as follows: maximum depth of the tree (max_depth) = 6; learning rate (eta) = 0.01; minimum sum of instance weight needed in a child (min_child_weight) = 1; subsample ratio of the training instance (subsample) = 1; fraction of features (columns) to use (colsample_bytree) = 1. The ratio between the training dataset and the test dataset was set to 0.8/0.2 in this study.

5.1. Optimal Time Window Size and Data Balance

We divided the audio signals collected from the test line into segment sequences with different time windows. Figure 6 presents the calculated Shannon entropies under different time window sizes. The Shannon entropy maintains a relatively stable state when the time window size increases from 0.1 (10^{-1}) to 1.58 ($10^{0.2}$) s, after which it decreases dramatically. When the time window size is 1.58 s, the Shannon entropy reached its maximum value. According to the maximum Shannon entropy hypothesis, the optimal time window size is 1.58 s. However, we maintained a relatively small window in our study to avoid a situation where one window contains different vehicle interior noise events. Therefore, we set the time window size to 1 s.
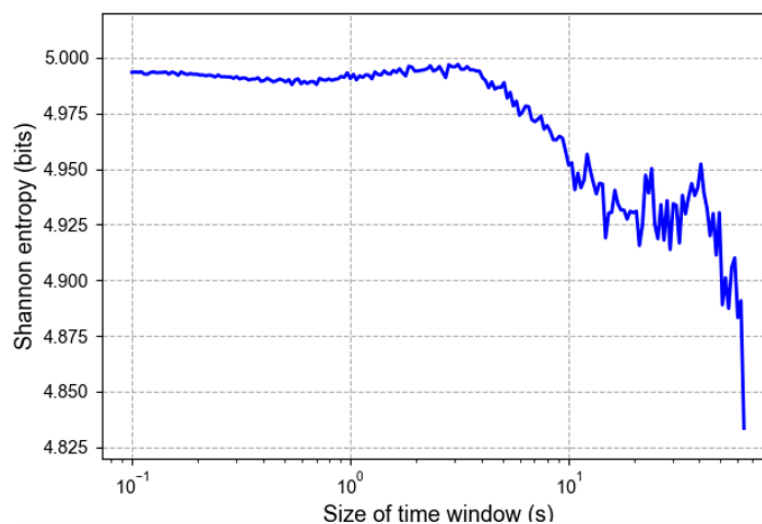

Figure 6. Entropy at different time window sizes.

We increased the proportion of four minority classes to the same as 'Broadcast' with SMOTE. The performance of the multi-classification model using balanced or unbalanced training data was compared. Table 2 reports the comparison results from the perspective of precision, recall, and F1 score. 'Support' in this table means the total number of occurrences in each category. Data balance increased the precision of 'Broadcast' and decreased its recall. In contrast, it decreased the precision and increased the recall of minority classes, namely 'Beep', 'Rumble', 'Squeal', and 'Other noises'. Meanwhile, F1 scores presented a slight drop after the data balance, except for the classes of 'Beep' and 'Squeal'.

Table 2. Classification reports of test results.

Classes	The Model Trained with Unbalanced Data			The Model Trained with Balanced Training Data			
	Precision	Recall	F1 score	Precision	Recall	F1 score	Support
Other noises	0.94	0.94	0.94	0.87	0.95	0.91	3671
Broadcast	0.96	0.98	0.97	0.98	0.92	0.95	11,274
Squeal	0.95	0.97	0.96	0.86	1.00	0.92	444
Rumble	0.95	0.89	0.92	0.82	0.97	0.89	466
Beep	0.95	0.73	0.83	0.70	0.87	0.78	834

We also employed confusion matrices to describe the performance before and after the training data were balanced, as shown in Figure 7. These matrices provide insights into the errors by the classification model and distinguish the types of errors. For instance, the matrices imply that 'Squeal' is commonly mislabeled as 'Broadcast', and 'Rumble' is mislabeled as 'Other noises'. One can also notice that the data balance improves the identification of the performance of minority classes such as 'Beep', 'Rumble', and 'Squeal'. 'Squeal' and 'Rumble' have a strong relationship with vehicle-track conditions, which is a major concern in our research. It is therefore desirable to detect all 'Squeal' and 'Rumble' events. Therefore, we balanced the training dataset via SMOTE to improve the recall of 'Squeal' and 'Rumble', despite the slight decrease in precision.
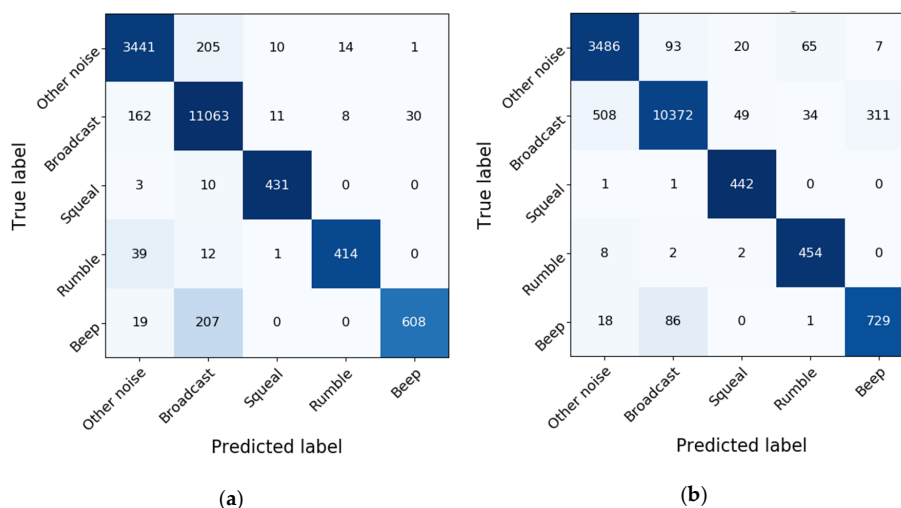

(a)

(b)

Figure 7. Confusion matrices of test results: (**a**) Model trained with unbalanced data; (**b**) Model trained with balanced data.

5.2. Feature Selection Based on the Importance Score

The importance was calculated explicitly for each feature by using the inbuilt feature importance property of XGBoost algorithm. The scores for features indicate how useful they were in the construction of the model and allows features to be ranked and compared with each other. Besides, a mutual information-based feature selection method is also used to verify the results of the importance-based method. In contrast to the importance score, the calculation of mutual information does not depend on the classifiers, but only considers the statistical characteristics of the input features and target variables.

In our classification model, 45 initial features were considered. Figure 8a shows the feature importance scores calculated by gain [45]. The importance scores of different features vary greatly, ranging from 0 to 378. The spectral centroid, denoted as f4, ranks first. In contrast, the importance score of f2, root mean square (RMS) of segments, equals zero, which means that it was not used during the training process. Figure 8a also shows that the low-order features and first 12 MFCCs are essential in the classification task. The results of the feature importance analysis indicate that the

contribution of different features to the model varies greatly. Thus, feature selection is necessary to improve the performance of the model and speed of calculations. Figure 8c shows the results for 45 features calculated by the mutual information-based method. The mutual information of these features has a similar trend with that of importance score. However, the importance scores of some features are very different from their mutual information value. For example, the importance score of feature f2 is 0, but its mutual information ranks fifth among all of the 45 features. The reason is that the mutual information only considering the features and target variables cannot reflect whether the features were engaged in the establishment of the classification model.
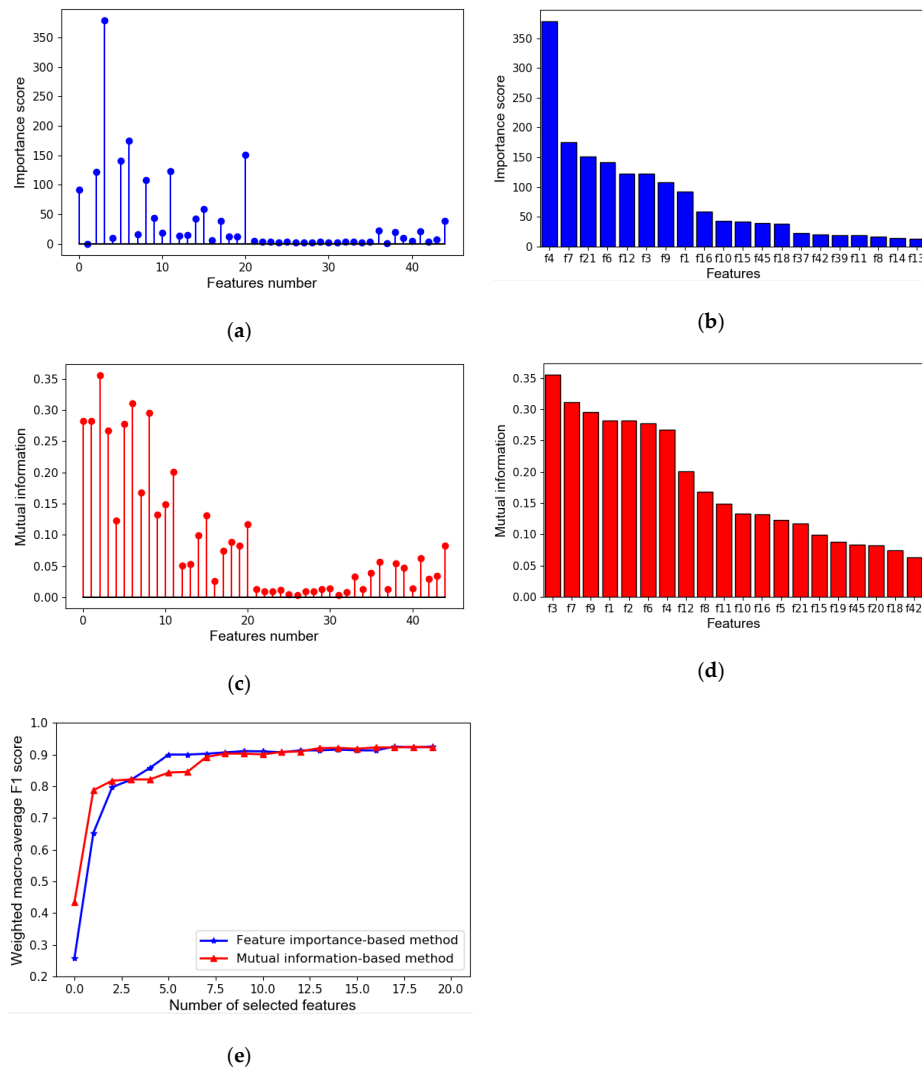

Figure 8. Illustration of feature selection based on different methods: (**a**) importance score of all the features; (**b**) importance score of the top 20 features; (**c**) mutual information of all the features; (**d**) mutual information of the top 20 features; (**e**) comparison of results of the two feature selection methods.

First, all 45 features were sorted in descending order of importance and mutual information, respectively. Figure 8b,d show the histograms of the top 20 features in descending order of the importance score and mutual information independently. We then constructed 20 feature sets incrementally with top 1, top 2, . . . , and top 20 features. Furthermore, the classification results with different features sets were compared, as shown in Figure 8e. There, the weighted macro average F1

score, $F1_{wm}$, was used to evaluate the performance of the multi-classification model, and it can be defined as follow:

$$F1_{wm} = \frac{\sum_{i=1}^{N} F1_i \times w_i}{N}, \tag{14}$$

where N is the total number of classes, in this study $N = 5$; $F1_i$ is the F1 score of the ith class; and w_i is the weight of the ith class and there is $\sum_{i=1}^{N} w_i = N$. Because this study mainly focuses on 'Squeal' and 'Rumble' we set both their weights to 1.3, and the weights of 'Other noises', 'Beep', and 'Broadcast', to 0.8. The value of $F1_{wm}$ varies from 0 to 1. The closer the weight is to 1, the better the model performs. The red line in Figure 8e corresponds to the classification results of 20 feature sets constructed by the mutual information-based feature selection method, and the blue line corresponds to that by the feature importance-based method. The results in Figure 8e show that $F1_{wm}$ by both feature selection methods increased rapidly when the feature set expanded from the top 1 to the top 8 features. Afterward, $F1_{wm}$ remained stable. The comparison of the results of the two methods indicates that the mutual information-based method performed better than the importance-based one when the number of selected features was less than 4. However, when the feature set expanded from the top 4 to the top 11, the importance-based method performed better. Then, the continuous increase in the number of the features selected causes no obvious difference between the performances of the two methods. According to the analysis, the set with the top 10 features selected by the importance-based method was employed in this study, the $F1_{wm}$ of which reached 0.91.

5.3. Comparisons with Other Methods

To validate the performance and execution speed of the XGBoost-based classifier used in our study, we conducted a comparison with other commonly used classifiers, including the K-nearest neighbors, decision trees, random forest, gradient boost, extra trees, AdaBoost, and artificial neural network (ANN) classifiers. This study ran all classifiers on the same computer and with the same training and testing data set. Table 3 shows the comparison results of $F1_{wm}$ and running time. The $F1_{wm}$ value of the gradient boost ranked first at 0.925. However, training and testing the gradient boost classifier also consumed the longest running time, 340.31 s, which was approximately 22 times longer than the time needed by the XGBoost classifier. In contrast, the K-nearest Neighbors presented the fastest computing speed and one of the lowest $F1_{wm}$. Besides, the accuracy and precision of different models are provided in Table 3. The accuracy and precision share a similar trend with $F1_{wm}$. The comparison with other classifiers depicts that the XGBoost model shows a good performance in accuracy and execution speed.

Table 3. Comparisons between XGBoost and other classifiers.

Classifier	$F1_{wm}$	Accuracy	Precision	Running Time (s)
XGBoost	0.923	0.96	0.95	15.06
K-nearest Neighbours	0.704	0.84	0.72	**2.51**
Decision Trees	0.851	0.91	0.92	3.12
Random Forest	0.923	0.96	0.94	77.88
Gradient Boost	**0.925**	0.96	0.94	340.31
AdaBoost	0.651	0.77	0.64	67.70
ANN	0.880	0.93	0.94	173.22

5.4. Case Studies to Extend the Model Application Scenarios

In this paper, we provided two case studies to extend the application scenarios. First, we conducted a statistical analysis to investigate the relationship between the vehicle interior noises and the dynamic responses of the car body with multi-source data collected by smartphones. After that, we used the proposed multi-classification model to detect abnormal interior noise events and evaluate the effect of

rail grinding for guiding the implementation of maintenance work. Figure 9 illustrates the schematics of both case studies in this work.
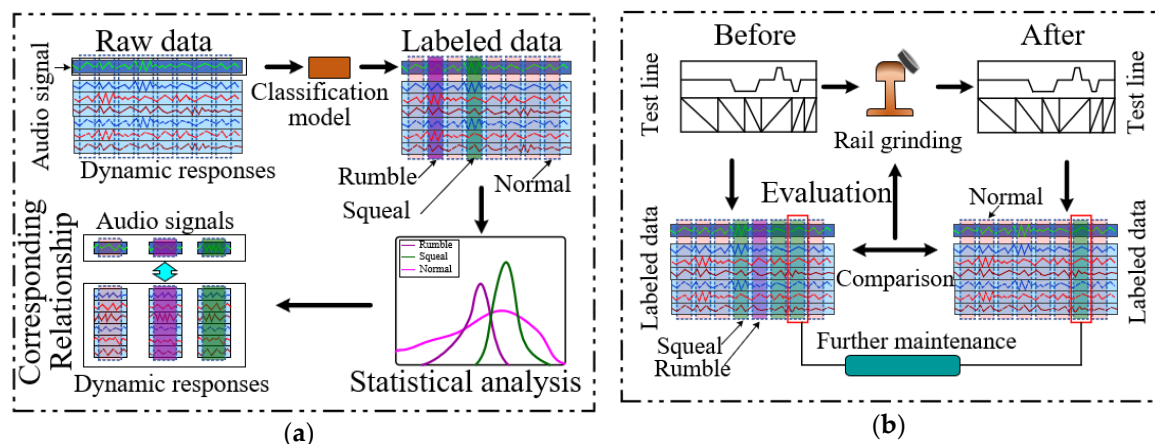

Figure 9. Schematics for case studies: (**a**) statistical analysis of vehicle interior noise and dynamic responses; (**b**) abnormal events detection and rail grinding effect evaluation using the XGBoost multi-classification model.

In the first case study, about 10 h of onboard monitoring data collected by smartphones were used. As shown in Figure 9a, the audio signals of the vehicle interior noise were fed into the multi-classification model established in this work. According to the classification results, the raw data were labeled into three categorizations: 'Squeal', 'Rumble', and 'Normal'. 'Normal' contained all other events except for 'Squeal' and 'Rumble' events. Then, statistical analyses for the dynamic responses corresponding to different vehicle interior noise were performed. This case study aimed to investigate the causes of the abnormal noise events and find out the solutions through the statistical analysis results.

For 'Squeal', 'Rumble', and 'Normal', the probability distribution curves of running speed (v) and vertical acceleration (a_v) of the car body are presented in Figure 9a,b, respectively. The vehicle speed v used here was not measured directly but obtained by the first-order integration of the longitudinal acceleration a_l [47], which can be written as follows:

$$v = \int_0^t a_l dt + v_0, \tag{15}$$

where t denotes the time; v_0 is the initial velocity. Since the integration begins when the subway train starts, v_0 equals to 0. The probability distribution curves in Figure 10a shows that 'Squeal' usually occurs at higher running speed compared with 'Normal' and 'Rumble'. This also suggests that we can reduce the occurrence of 'Squeal' by adjusting the operating speed of the train. In contrast, 'Rumble' occurs at a slower speed and higher vertical vibration level compared to 'Squeal', as shown in Figure 10b. This phenomenon implies that the occurrence of 'Rumble' is related to the resonance of the car body, which may be avoided by optimizing the structure of the car body.

The schematic of the second case study is presented in Figure 9b. The test interval selected in this study was between two adjacent stations with a length of 1631 m. The track alignment of the test interval is presented in the upper plot of Figure 11a. There are three curves in the test interval, the radii of which are 1200 m, 800 m, and 800 m. This case study aimed to test the capacity of this model for identifying abnormal noise events, evaluating the effect of rail grinding, and providing information relevant to designing a future maintenance plan.
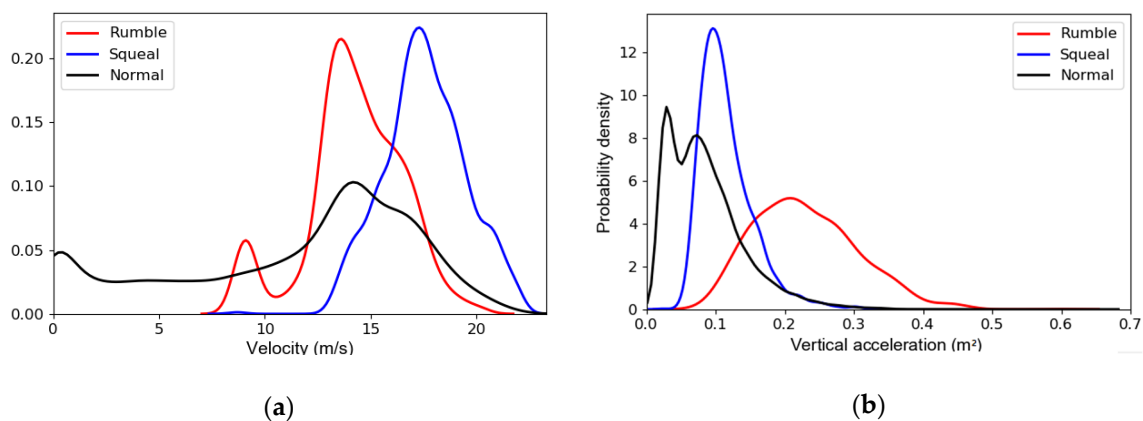
Figure 10. Statistical analysis of vehicle interior noise and dynamic responses: (**a**) The probability distribution curves of running speed (*v*); (**b**) The probability distribution curves of vertical acceleration (a_v).
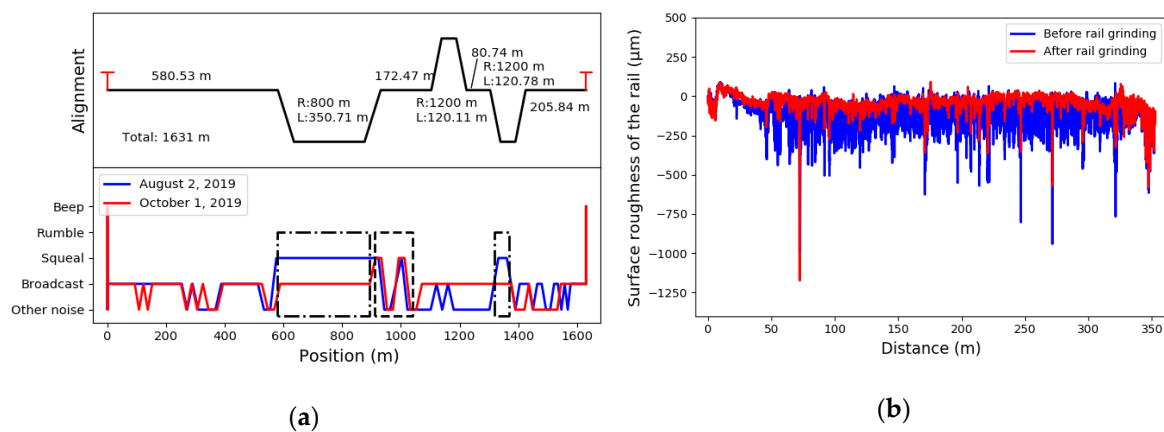

Figure 11. Abnormal events detection and rail grinding effect evaluation using the XGBoost multi-classification model: (**a**) track alignments of the test section and the identification results before and after rail grinding; (**b**) the surface roughness of the rail before and after rail grinding.

The authors first collected multi-source data with the onboard smartphone on 2 August 2019. The results of the multi-classification model are depicted in the lower plot of Figure 11a with a blue line. The results indicate that 'Squeal' occurred in the positions from 580 to 890 m, 910 to 1040 m, and 1320 to 1370 m. It can be seen that the figure the sections where 'Squeal' occurs have a high overlap ratio with the curve sections, especially the curve section with a radius of 800 m. According to the classification results and design information, we can make a preliminary conclusion that the sharp curves are the main causes of 'Squeal'. The results also indicate the need for rail grinding or other corresponding maintenance measures.

Then, a scheduled rail grinding of the test interval was done on 21 August 2019. The surface roughness of the rail before and after rail grinding presented in Figure 11b indicates that rail grinding reduced the roughness of the rail surface effectively. Since reducing the rail roughness, that is, the unevenness on the tread of the rail benefits improving the rail-wheel contact relationship, rail grinding is a common measure for eliminating the abnormal noise and vibration of subway trains.

Another onboard test was conducted on 1 October 2019, to verify the effects of the maintenance work. The corresponding classification results after the rail grinding are displayed in red in the lower plot of Figure 11a. It can be seen that after rail grinding, the 'Squeal' was eliminated at 580–890 m and 1320–1370 m. However, the 'Squeal' at 910–1040 m remained. The results illustrate that rail grinding eliminated 'Squeal' at circular curves effectively. Nevertheless, it showed no apparent effect on the occurrences at transition curves and straight-line sections, which shows that there exist some other

factors that lead to 'Squeal' in these sections. Thus, future maintenance work should focus on the section from 910 to 1040 m. This case study demonstrates the potential of applying the proposed multi-classification model in evaluating the effect of rail grinding and providing more information about the track conditions to making a further rail maintenance plan.

6. Conclusions

This study proposed a vehicle interior noise multi-classification model based on the XGBoost method and onboard smartphone data. By considering the Shannon entropy, a 1-second time window was selected to perform the data segmentation task. The comparison between the performances before and after the training data was balanced demonstrated that data balancing can promote the recall of minority classes but decrease the precision of their results. Feature importance analysis results show that features calculated from the spectrum of the Fourier transform and the first 12 MFCCs are the most essential among all features. By comparing and analyzing the results of importance-based and mutual information-based methods, this study selected the top 10 features in importance score to form the features set, whose $F1_{wm}$ reached 0.91. Then, the comparison between the XGBoost and other commonly used classifiers showed that the proposed XGBoost-based classification model presents a faster computing speed while maintaining a good performance. The case studies verified that the proposed multi-classification model has the potential to investigate the correlation between abnormal vehicle interior noise and dynamic responses of the train. Moreover, the capacity of the model to monitor abnormal noise events and evaluate the effect of rail grinding was also proved.

There are a few directions for future research. A more detailed classification of vehicle interior noise could be developed based on specific track-vehicle conditions so that this model would be suitable for general cases. Furthermore, more experiments are needed to explain the performance among different vehicles and track slabs. Another interesting option is to investigate the relationship between abnormal noise and wheel-rail contact conditions. Furthermore, the authors intend to set up a data collection system with high-quality sensors for more accurate and reliable data.

Author Contributions: Conceptualization, P.W. and Q.H.; data curation, Y.W. and Q.H.; formal analysis, Y.W.; methodology, Y.W. and Q.W.; validation, Y.W. and Z.C.; writing—original draft preparation, Y.W.; writing—review and editing, Y.W. and Q.H. All authors have read and agreed to the published version of the manuscript.

References

1. China Urban Rail Transit Association. *Urban Rail Transit 2018 Annual Statistical Report*; China Urban Rail Transit Association: Beijing, China, 2019.
2. Atmaja, B.; Puabdillah, M.; Farid, M.; Asmoro, W. Prediction and simulation of internal train noise resulted by different speed and air conditioning unit. *J. Phys. Conf. Ser.* **2018**, *1075*, 012038. [CrossRef]
3. Zhang, J.; Xiao, X.; Sheng, X.; Li, Z.; Jin, X. A Systematic Approach to Identify Sources of Abnormal Interior Noise for a High-Speed Train. *Shock Vib.* **2018**, *2018*. [CrossRef]
4. Talotte, C. Aerodynamic noise: A critical survey. *J. Sound Vib.* **2000**, *231*, 549–562. [CrossRef]
5. Han, J.; Xiao, X.; Wu, Y.; Wen, Z.; Zhao, G. Effect of rail corrugation on metro interior noise and its control. *Appl. Acoust.* **2018**, *130*, 63–70. [CrossRef]
6. Wu, B.; Chen, G.; Lv, J.; Zhu, Q.; Kang, X. Generation mechanism and remedy method of rail corrugation at a sharp curved metro track with Vanguard fasteners. *J. Low Freq. Noise Vib. Act. Control.* **2019**. [CrossRef]
7. Li, L.; Thompson, D.; Xie, Y.; Zhu, Q.; Luo, Y.; Lei, Z. Influence of rail fastener stiffness on railway vehicle interior noise. *Appl. Acoust.* **2019**, *145*, 69–81. [CrossRef]

8.   Meehan, P.A.; Liu, X. Modelling and mitigation of wheel squeal noise amplitude. *J. Sound Vib.* **2018**, *413*, 144–158. [CrossRef]

9.   Zhang, J.; Han, G.; Xiao, X.; Wang, R.; Zhao, Y.; Jin, X. Influence of Wheel Polygonal Wear on Interior Noise of High-Speed Trains. In *China's High-Speed Rail Technology*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 373–401.

10.   Fink, O.; Zio, E.; Weidmann, U. Predicting time series of railway speed restrictions with time-dependent machine learning techniques. *Expert Syst. Appl.* **2013**, *40*, 6033–6040. [CrossRef]

11.   Sun, Y.; Zhao, Y. Characteristics of Interior Noise in MonoRail and Noise Control. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*; Institute of Noise Control Engineering: Chicago, IL, USA, 2018; Volume 258, pp. 1461–1467.

12.   Hu, K.; Wang, Y.; Guo, H.; Chen, H. Sound quality evaluation and optimization for interior noise of rail vehicle. *Adv. Mech. Eng.* **2014**, *6*, 820875. [CrossRef]

13.   Kurzweil, L.G. Prediction and control of noise from railway bridges and tracked transit elevated structures. *J. Sound Vib.* **1977**, *51*, 419–439. [CrossRef]

14.   Zhang, J.; Xiao, X.; Sheng, X.; Li, Z. Sound Source Localisation for a High-Speed Train and Its Transfer Path to Interior Noise. *Chin. J. Mech. Eng.* **2019**, *32*, 59. [CrossRef]

15.   Zhang, J.; Xiao, X.; Sheng, X.; Zhang, C.; Wang, R.; Jin, X. SEA and contribution analysis for interior noise of a high speed train. *Appl. Acoust.* **2016**, *112*, 158–170. [CrossRef]

16.   Franzoni, L.; Rouse, J.; Duvall, T. A broadband energy-based boundary element method for predicting vehicle interior noise. *J. Acoust. Soc. Am.* **2004**, *115*, 2538. [CrossRef]

17.   Wu, D.; Ge, J.M. Analysis of the Influence of Racks on High Speed Train Interior Noise Using Finite Element Method. *Appl. Mech. Mater.* **2014**, *675*, 257–260. [CrossRef]

18.   Ghofrani, F.; He, Q.; Goverde, R.M.; Liu, X. Recent applications of big data analytics in railway transportation systems: A survey. *Transp. Res. Part. C Emerg. Technol.* **2018**, *90*, 226–246. [CrossRef]

19.   Toque, F.; Come, E.; Oukhellou, L.; Trepanier, M. Short-Term Multi-Step Ahead Forecasting of Railway Passenger Flows During Special Events With Machine Learning Methods. In Proceedings of the CASPT 2018, Conference on Advanced Systems in Public Transport and TransitData 2018, Brisbane, Australia, 23–25 July 2018; p. 15.

20.   Cui, Y.; Martin, U.; Zhao, W. Calibration of disturbance parameters in railway operational simulation based on reinforcement learning. *J. Rail Transp. Plan. Manag.* **2016**, *6*, 1–12. [CrossRef]

21.   Ghofrani, F.; Pathak, A.; Mohammadi, R.; Aref, A.; He, Q. Predicting rail defect frequency: An integrated approach using fatigue modeling and data analytics. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *35*, 101–115. [CrossRef]

22.   Mohammadi, R.; He, Q.; Ghofrani, F.; Pathak, A.; Aref, A. Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects. *Transp. Res. Part C Emerg. Technol.* **2019**, *102*, 153–172. [CrossRef]

23.   Ghofrani, F.; He, Q.; Mohammadi, R.; Pathak, A.; Aref, A. Bayesian Survival Approach to Analyzing the Risk of Recurrent Rail Defects. *Transp. Res. Rec.* **2019**, *2673*, 281–293. [CrossRef]

24.   Li, H.; Parikh, D.; He, Q.; Qian, B.; Li, Z.; Fang, D.; Hampapur, A. Improving rail network velocity: A machine learning approach to predictive maintenance. *Transp. Res. Part C Emerg. Technol.* **2014**, *45*, 17–26. [CrossRef]

25.   He, Q.; Li, H.; Bhattacharjya, D.; Parikh, D.P.; Hampapur, A. Track geometry defect rectification based on track deterioration modelling and derailment risk assessment. *J. Oper. Res. Soc.* **2015**, *66*, 392–404. [CrossRef]

26.   Li, Z.; He, Q. Prediction of railcar remaining useful life by multiple data source fusion. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2226–2235. [CrossRef]

27.   Verhaegh, W.; Verhaegh, W.; Aarts, E.; Korst, J. *Algorithms in Ambient Intelligence*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2004; Volume 2.

28.   Mato-Méndez, F.J.; Sobreira-Seoane, M.A. Blind separation to improve classification of traffic noise. *Appl. Acoust.* **2011**, *72*, 590–598. [CrossRef]

29.   Sobreira-Seoane, M.A.; Rodriguez Molares, A.; Alba Castro, J.L. Automatic classification of traffic noise. *J. Acoust. Soc. Am.* **2008**, *123*, 3823. [CrossRef]

30.   Paulraj, P.; Melvin, A.A.; Sazali, Y. Car Cabin Interior Noise Classification Using Temporal Composite Features and Probabilistic Neural Network Model. *Appl. Mech. Mater.* **2014**, *471*, 64–68. [CrossRef]

31.   Wang, Y.; Wang, P.; Wang, X.; Liu, X. Position synchronization for track geometry inspection data via big-data fusion and incremental learning. *Transp. Res. Part C Emerg. Technol.* **2018**, *93*, 544–565. [CrossRef]

32. Cho, C.J.; Park, Y.; Ku, B.; Ko, H. An implementation of environment recognition for enhancement of advanced video based railway inspection car detection modules. *Sci. Adv. Mater.* **2018**, *10*, 496–500. [CrossRef]

33. Yin, J.; Zhao, W. Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach. *Eng. Appl. Artif. Intell.* **2016**, *56*, 250–259. [CrossRef]

34. Li, C.; Luo, S.; Cole, C.; Spiryagin, M. An overview: Modern techniques for railway vehicle on-board health monitoring systems. *Veh. Syst. Dyn.* **2017**, *55*, 1045–1070. [CrossRef]

35. Tsunashima, H.; Naganuma, Y.; Matsumoto, A.; Mizuma, T.; Mori, H. Condition monitoring of railway track using in-service vehicle. *Reliab. Saf. Railw.* **2012**, *12*, 334–356.

36. Wang, P.; Wang, Y.; Wang, L.; Chen, R.; Xiao, J. Measurement of Carbody Vibration in Urban Rail Transit Using Smartphones. In Proceedings of the Transportation Research Board 96th Annual Meeting, Washington, DC, USA, 8–12 January 2017.

37. Ghose, A.; Biswas, P.; Bhaumik, C.; Sharma, M.; Pal, A.; Jha, A. Road condition monitoring and alert application: Using in-vehicle Smartphone as Internet-connected sensor. In Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, Lugano, Switzerland, 19–23 March 2012; pp. 489–491.

38. Han, W.; Chan, C.F.; Choy, C.S.; Pun, K.P. An efficient MFCC extraction method in speech recognition. In Proceedings of the 2006 IEEE International Symposium on Circuits and Systems, Island of Kos, Greece, 21–24 May 2006.

39. Banos, O.; Galvez, J.-M.; Damas, M.; Pomares, H.; Rojas, I. Window Size Impact in Human Activity Recognition. *Sensors* **2014**, *14*, 6474–6499. [CrossRef]

40. Zhang, X.; Feng, N.; Wang, Y.; Shen, Y. Acoustic emission detection of rail defect based on wavelet transform and Shannon entropy. *J. Sound Vib.* **2015**, *339*, 419–432. [CrossRef]

41. Sun, Y.; Wong, A.K.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]

42. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer Science + Business Media: Berlin/Heidelberg, Germany, 2006.

43. Martinez, J.; Perez, H.; Escamilla, E.; Suzuki, M.M. Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques. In Proceedings of the CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers, Puebla, Mexico, 27–29 February 2012; pp. 248–251.

44. Lei, S. A feature selection method based on information gain and genetic algorithm. In Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering, Hangzhou, China, 23–25 March 2012; Volume 2, pp. 355–358.

45. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.

46. Omar, K. XGBoost and LGBM for Porto Seguro's Kaggle Challenge: A Comparison. 2018. Available online: https://pub.tik.ee.ethz.ch/students/2017-HS-SA-2017-98.pdf (accessed on 7 July 2019).

47. Wang, Y.; Cong, J.; Tang, H.; Liu, X.; Gao, T.; Wang, P. A Data Fusion Approach for Speed Estimation and Location Calibration of a Metro Train. in Underground Environment Based on Low-Cost Sensors in Smartphones. In Proceedings of the Transportation Research Board 98th Annual Meeting, Washington, DC, USA, 13–17 January 2019.