Safety Emergency Science

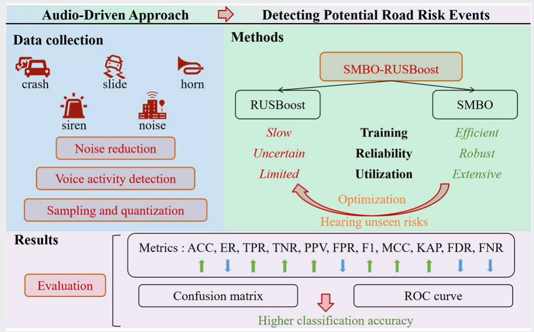
Research Article

Hearing unseen risks: An advanced audio-driven approach for detecting potential road risk events

Jiming Xie[1,2,†], Yan Zhang[1,†], Bijun Wang[1,†], Han Han[3,†], Yaqin Qin[1,✉], Yulan Xia[4,✉]

Cite this article: Xie JM, Zhang Y, Wang BJ, et al. *Safety Emergency Science*, 2025, **1**(3): 9590018. https://doi.org/10.26599/SES.2025.9590018

ABSTRACT: Traditional road monitoring methods, including radar, light detection and ranging (LiDAR), and video surveillance, face challenges in accurately and comprehensively identifying unexpected events and potential risks owing to high costs, low accuracy, and environmental complexities. This study proposes an audio-based monitoring method for detecting abnormal road events aimed at uncovering hidden traffic risks. Audio samples collected from various sources, such as FreeSound and Zapslat, are preprocessed using noise reduction, detection, integration, and digital signal conversion via sample quantization. An audio recognition classification model based on the RUSBoost algorithm is constructed and optimized using sequential model-based optimization (SMBO). This optimization addresses issues related to imbalanced samples and features. Optimization increased the accuracy from 68.9% to 96.5% and reduced the error rate to 3.5%. The true positive rate (TPR), true negative rate (TNR), and positive predictive value (PPV) also showed significant improvements, along with notable enhancements in the false positive rate (FPR), F1 score, matthews correlation coefficient (MCC), and kappa (KAP). For key audio events such as slides and crashes, TPR and PPV exceeded 93.2%. The SMBO-RUSBoost model distinguished traffic events from noise by extracting multidimensional features from audio data corresponding to various traffic events. Its classification performance provides a solid foundation for traffic risk assessment and decision-making. The model offers a novel approach for improving road safety.



KEYWORDS: audio-based monitoring; traffic risk detection; RUSBoost; sequential model-based optimization (SMBO); road safety

1 Introduction

In recent years, economic progress has led to a rapid development in road traffic, which has inevitably been accompanied by an increase in frequent accidents. When accidents occur, the stability of traffic flow is broken, leading to traffic jams and, in some cases, secondary accidents. In such conditions, secondary accidents can become even more hazardous, as the disturbed traffic environment increases the likelihood of subsequent collisions (Tedesco et al., 1994).

In this context, achieving real-time monitoring of road conditions would enable prompt emergency responses, helping to avoid congestion and secondary accidents, from a management perspective. At the planning level, early identification of potential safety hazards could assist relevant authorities in improving the road environment. Consequently, developing effective road-monitoring systems has become a central issue in reducing traffic accidents.

Previously, road monitoring approaches were based on radar, light detection and ranging (LiDAR), radio-frequency identification (RFID), and laser-based systems. However, their high cost and low precision have limited their application to specific fields, such as speed monitoring (Boyarchikov and Martinec, 2024; Cui et al., 2020; Khan et al., 2019; Kim et al., 2016; Liu et al., 2022). These modalities are often constrained by line-of-sight and high deployment costs. For a more comprehensive understanding of traffic conditions, video recognition is included in the monitoring system (García-González et al., 2021; Kotov and Pospelov, 2017; Pramanik et al., 2021; Rathore et al., 2021). However, vehicle detection, tracking, and data correlation present significant challenges owing to factors such as partial or complete object occlusion, camera instability, and varying weather conditions (e.g., rain, snow, and high winds). These challenges not only reduce the detection accuracy of video-based surveillance but also increase the likelihood of overlooking critical traffic events, thereby exacerbating "invisible traffic risks". Consequently, relying

† Jiming Xie, Yan Zhang, Bijun Wang, and Han Han contributed equally to this work.

[1] Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming, China. [2] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China. [3] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China. [4] Department of Traffic Engineering, University of Shanghai for Science and Technology, Shanghai, China.

✉ Corresponding authors. E-mail: Y. Qin, qinyaqin@kust.edu.cn; Y. Xia yulan_xia@163.com

Received: May 20, 2025; Revised: November 20, 2025; Accepted: December 16, 2025

solely on video detection may fail to comprehensively and accurately identify potential risks in complex traffic environments (Datondji et al., 2016). To address these challenges, Azimjonov and Özmen (2021) proposed a robust, real-time tracking algorithm for more effective and efficient video analysis. With the advent of the Big Data era, various data sources, such as demographic data, weather reports, road geometry characteristics, and crash data, have been extensively used in traffic operations, safety management, and research (Thabit et al., 2024). However, monitoring and reporting abnormal events remains challenging.

Environmental sound classification techniques have been widely applied in diverse scenarios because of their low cost and high quality (Ibili et al., 2022; Iglesias-Merchan et al., 2021; Marciniuk and Kostek, 2022; Pascale et al., 2023). Based on noise monitoring and traffic speed data, Lan and Cai (2021) proposed a method for dynamically updating traffic noise maps in large regions. Using a two-stage structural equation modeling–artificial neural network (SEM-ANN) approach, Prasad Das et al. (2021) evaluated the effects of traffic noise exposure. However, sound classification has yet to be implemented for road monitoring. Therefore, we propose a model that can identify emergencies using special event audio data. Owing to its low cost and fast data transmission, the model can also analyze potential safety risks by detecting multiple and regular events such as car horns at a particular location.

To monitor roadway emergencies based on audio data, we first collected five types of sound data from open-source datasets: crashes, slides, horns, sirens, and city noise. Next, we converted the audio features into digital forms and extracted relevant characteristics for data preprocessing. Finally, we used RUSBoost, an integrated machine-learning algorithm, to build the model. To further improve the prediction accuracy, we optimized hyperparameters using sequential model-based optimization (SMBO) algorithms.

The key contributions of this study are as follows:

• Augmentation of existing systems with auditory sensors to improve the accuracy and reliability of road monitoring. To the best of our knowledge, this study builds a novel end-to-end audio-driven framework integrating real-time risk voice recognition applied to road monitoring.

• An RUSBoost model optimized using SMBO to address challenges related to uneven data distribution and unclear feature representation. By fine-tuning the hyperparameters, the optimized model significantly improved the accuracy in event recognition and audio sample classification compared to traditional models.

In general, an audio-based method for roadway emergency monitoring can detect abnormal events and potential road risks, improving safety and stability. The development of intelligent systems capable of extracting certain events from audio data has the potential to contribute significantly to smart city initiatives.

2 Dataset

Although several open-source environmental sound datasets exist, specialized audio events such as crashes, slides, horns, and sirens are not available. In this study, we expand the existing ESC-50 (Piczak, 2015) sound dataset by incorporating these specific audio clips. These events, as shown in Table 1, are collected from different sources, including FreeSound (https://freesound.org/) and Zapslat (https://www.zapsplat.com/).

This section describes the process of transforming the collected sound data into identifiable digital forms. First, to reduce the noise generated during acquisition and transmission, we employed an improved wavelet threshold denoising algorithm based on ensemble empirical mode decomposition (EEMD). Next, to better distinguish audio from noise, rather than solely using energy characteristics, we employed a dual threshold for voice activity detection. By integrating energy with the zero-crossing rate (ZCR), we extracted relevant audio clips for the targeted special events. Finally, during sampling and quantization, analog audio signals are converted into digital signals, represented as discrete values for time and amplitude.

2.1 Noise reduction

Noise is inevitably introduced into audio samples due to natural or artificial factors during data collection and processing. Noise reduction aims to retain the original signal of the measured data while minimizing noise. For nonstationary audio samples, we applied an improved wavelet threshold denoising algorithm based on EEMD (Feng et al., 2010). By adding white noise to the original signal, EEMD skillfully solves the mode-mixing problem existing in traditional empirical mode decomposition (EMD). Conversely, with better local time–frequency characteristics, wavelet threshold noise reduction preserves detailed information, compensating for EEMD's limitations in handling complex noise signals. The noise reduction process follows these steps:

Step 1: Decompose the signal $X(t)$ using EEMD to obtain intrinsic mode function (IMF) components (Wu and Huang, 2009). The detailed process is shown in Fig. 1.

Step 2: Perform wavelet threshold denoising on the signaling components.

Step 3: Reconstruct the signal using the noise-reduced IMF components and residue to obtain a denoised output.

2.2 Voice activity detection

For long audio recordings, voice activity detection is essential for identifying the start and end points of crucial segments. This process not only reduces storage and processing demands but also eliminates noise interference from silent segments. We adopted a dual-threshold detection method to determine these endpoints (Liu and Tian, 2016). Combining short-term energy with ZCR can effectively enhance the detection accuracy.

The main difference between effective and noisy signals lies in their energy levels. Short-term energy directly reflects changes in audio signal energy and amplitude (Han and Wang, 2015). Generally, the energy of an effective signal segment is higher than that of noise. First, we considered separating speech segments from the background noise by setting a threshold for short-time energy, which can be defined as

Table 1 Audio samples of special traffic incidents

Serial number	Form	Sound type	Sample size
1	Crash	Front-end collision, side-end collision, rear-end collision, scrape, run over, rollover, crash, crash into fixtures	16
2	Slide	Dry, waterlogged, flooded, icy, slushy, oily road surface	6
3	Horn	Long beeps, short beeps, continuous beeps	6
4	Siren	General vehicles, fire engines, police vehicles, engineering rescue vehicles, ambulances	5
5	Noise	Engines, airplanes, footsteps, alarms, fireworks, helicopters, bells	7

$$\mathrm{E}_n = \sum_{m=-\infty}^{+\infty} (x(m)w(n-m))^2 \tag{1}$$

Here, $x(\cdot)$ is the input sampling series of the preprocessed signal, and $w(n-m)$ is the Hamming and Hann window function. m denotes the local index of sampling points within a single frame, whereas n denotes the global index of sampling points in the entire audio signal.

The ZCR of speech, denoted as Z_n, represents the number of zero crossings per 10-ms interval. It can be calculated using Eq. (2). Although ZCR is highly susceptible to 60-Hz hum and direct current (DC) offset, it remains a reliable indicator of silent clips in most cases (Rabiner and Sambur, 1975).

$$\mathrm{Z}_n = \sum_{m=-\infty}^{+\infty} |\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]w(n-m)| \tag{2}$$

where the operator sgn[·] represents the sign function.

Using short-term energy and ZCR, we implemented a dual-threshold method to detect signal endpoints twice (Fig. 2). The detailed steps are as follows.

Step 1: First, a higher threshold T_1 is selected for the short-time energy envelope. An initial approximation is made based on this assumption, considering the segment above the threshold as a valid sound clip (segment *AB*). A lower threshold T_2 is then determined to complete the first-level judgment. We searched outward from points *A* and *B* to find two points, *C* and *D*, where the short-time energy envelope intersects T_1. These points, *C* and *D*, are identified as the start and end of the audio segment, respectively, based on the dual-threshold method applied to the short-time energy (Rabiner, 1978; Rabiner and Sambur, 1975).

Step 2: Using the ZCR, we performed a search from point *C* to the left and from point *D* to the right to identify two points *E* and *F*, where ZCR falls below a threshold T_3. These points mark the start and end of the audio segment (Rabiner, 1978; Huang et al., 2001). The area between them is the effective area that must be extracted.

2.3 Audio reading and feature extraction

In the real world, sound exists as continuous analog values that effectively represent an infinite range of amplitudes (loudness) and frequencies. However, digital technology cannot process infinite values. When analog signals are digitized, their infinite continuous scales of analog signals must be reduced to a finite range of discrete bits and bytes through sampling and quantization (Waggoner, 2009).

Sampling determines discrete points or regions to be measured. The sampling rate refers to the number of sampling points per unit time. Generally, a higher sampling rate results in more

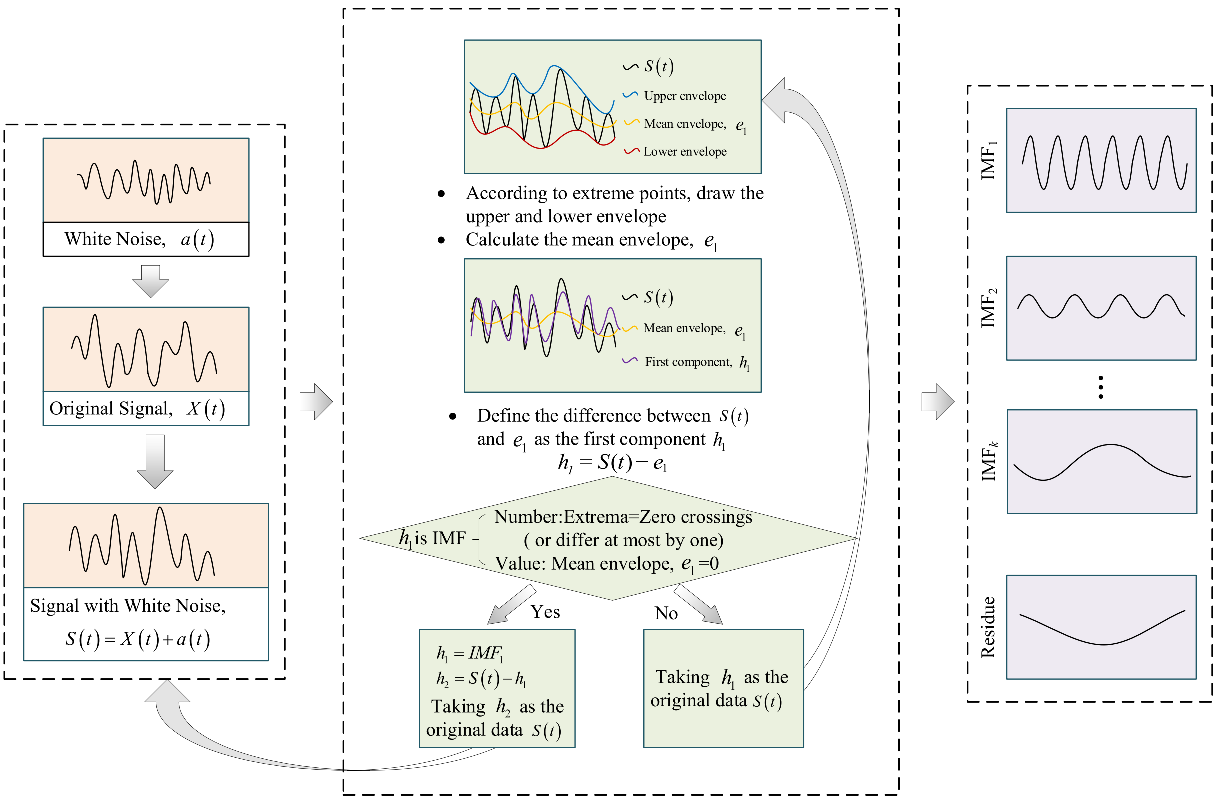


Fig. 1 Principle of the EEMD algorithm.

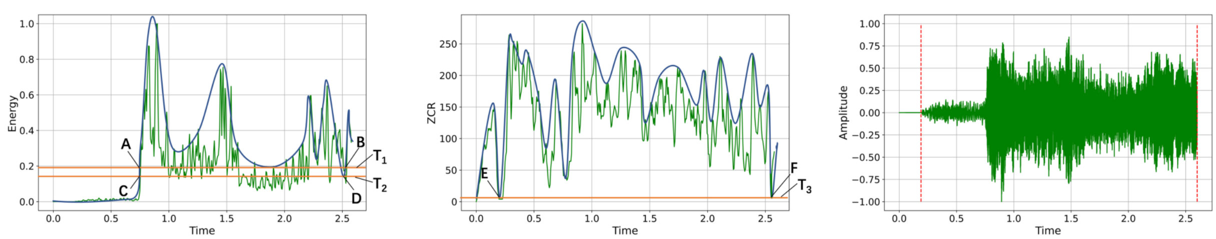


Fig. 2 A dual-threshold method to detect signal endpoints.

清華大学出版社 Tsinghua University Press | SciOpen

frequent signal measurements per unit time, improving sound quality. According to Nyquist–Shannon sampling theorem (Por et al., 2019), if the sampling rate is less than twice the frequency of the signal being sampled, aliasing may occur, affecting the reliability of the result. Therefore, to ensure reliability, the sampling rate must exceed twice the signal frequency. Considering both data accuracy and operation efficiency, we sampled the audio at 44,100 Hz.

As shown in Fig. 3, quantization defines the actual recorded values by digitizing the signal in the amplitude direction and recording the amplitude values of the sampled points. Finally, we applied a linear transformation on the amplitude values, mapping the data in the range [−1, 1] (Eq. (3)).

$$x^* = 2\frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1 \tag{3}$$

3 Methods

Classification accuracy is critical for identifying and monitoring abnormal road events. The AdaBoost algorithm is known for its strong learning ability and effective classification performance (Freund and Schapire, 1996). The core idea of AdaBoost is to train a base learner using an initial training dataset that comprises preprocessed audio samples. Then, we obtain classification results, including crashes and slides. The algorithm iteratively adjusts sample weights based on the classification results of the base learner, increasing the weights of misclassified samples and decreasing those of correctly classified samples. This iterative process refines the weak learner, ultimately enhancing it into a strong learner with higher prediction accuracy. Studies have shown that RUSBoost, being simpler and faster than AdaBoost, often achieves significantly better classification performance (Rahman and Zhu, 2023).

As an evolution model of AdaBoost, RUSBoost applies random undersampling (RUS), a technique that randomly removes examples from the majority class to balance the dataset (Seiffert et al., 2010). The model architecture of RUSBoost is illustrated in Fig. 4. Owing to the imbalance of samples and features, the RUSBoost model may not be able to classify accurately. In addition, RUSBoost suffers from a long training time and uncertainty in weak classifiers. To address these issues, we employed SMBO to automatically optimize the hyperparameters, enabling deeper exploration of hidden relationships between events and features.

Hyperparameters refer to the parameters of the training algorithm. Each model has different hyperparameters, such as learning rate, number of layers, and number of neurons in each layer of the neural network model. Optimal hyperparameter selection enables the algorithm to achieve optimal performance. Several hyperparametric optimization methods exist, including grid search, random search, and Bayesian optimization. Given that grid and random searches can be time-consuming and unsatisfactory, an autotuning strategy that automatically sets the hyperparameters to optimize performance is required (Wang et al., 2022).

The Bayesian optimization approach tracks past evaluation results to create a probabilistic model of the training objective function. By mapping hyperparameters to evaluation results, this method effectively searches the hyperparameter space. In addition, it offers high calculation efficiency, comprehensive information utilization, and robust problem-solving capabilities. SMBO, a simplified form of Bayesian optimization, prioritizes promising hyperparameters by minimizing expected loss while maximizing variance within a limited number of iterations (Wang et al., 2022).

If the unknown objective function is f, the SMBO algorithm can be expressed as

$$\xi * = \arg\min f(\xi), \quad \xi \in \Omega \tag{4}$$

Here, $f(\xi)$ represents a measure for evaluating input performance, i.e., root mean squared error (RMSE) and loss on the verification set. $\xi *$ indicates the hyperparametric combination that produces the minimum RMSE loss. ξ refers to the d-dimensional hyperparametric decision vector, and Ω represents the hyperparameter decision domain space. The specific process is as follows.

Step 1: Modeling is conducted using historical observations. Within the range of hyperparameters for the event recognition and classification model, t initialization sample points are randomly generated and denoted by $D_{1:t} = \{\xi_{1:t}, \eta_{1:t}\}$, where $\eta_t = f(\xi_t)$. A tree-structured Parzen estimator (TPE) is used to fit a surrogate probability model based on existing data.

$$p(\eta|\xi) = \frac{p(\xi|\eta)p(\eta)}{p(\xi)} \tag{5}$$

$$p(\xi|\eta) = \begin{cases} l(\xi), & \eta < \eta^* \\ g(\xi), & \eta > \eta^* \end{cases} \tag{6}$$

where $p(\xi|\eta) = \begin{cases} l(\xi), & \eta < \eta^* \\ g(\xi), & \eta > \eta^* \end{cases}$ denotes the distribution of hyperparameters associated with better performance $l(\xi)$ and poorer performance $g(\xi)$. γ (usually set to 0.25) corresponds to the proportion of samples with performance better than the threshold η^*. R denotes the set of all real numbers. Therefore,

$$p(\xi) = \int_R p(\xi|\eta)p(\eta)\mathrm{d}\eta = \gamma l(\xi) + (1-\gamma)g(\xi) \tag{7}$$

Step 2: Based on the prior distribution of the current probability proxy model, the expected improvement (EI) is used as the acquisition function to obtain the sampling points. The equations are as Eqs. (8) and (9):

$$\xi_{\text{new}} = \arg\max_{\xi} \mathrm{EI}_{\eta^*}(\xi) \tag{8}$$

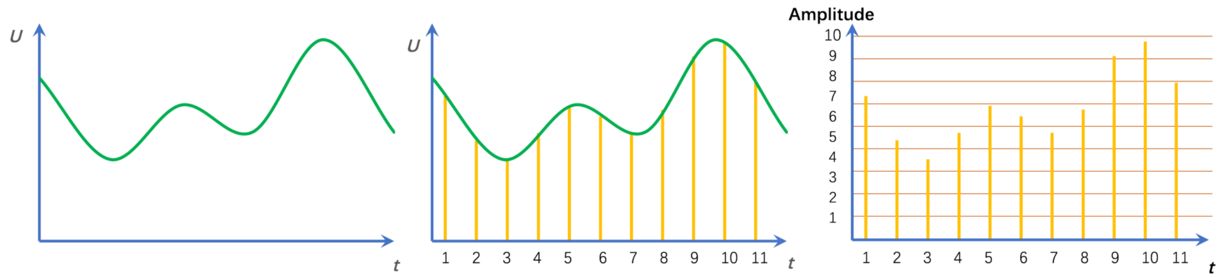


Fig. 3 Sampling and quantization.

Algorithm RUSBoost

Given: Set training set P of examples $(w_1, y_1), \ldots, (w_b, y_b)$ with minority class $y^r \in Y$, binary classification $|Y| = 2$

Weak learner, $WeakLearn$

Number of iterations, Q

Desired percentage of total instances to be represented by the monority class, N

1 Initialize the weight $V_1(i) = \frac{1}{b}$ for all i.

2 Do for iteration $q = 1, 2, \cdots, Q$

a Create temporary training dataset P_q' with distribution V_q' using random undersampling.

b Call $WeakLearn$, providing it with examples P_q' and their weights V_q'.

c Get back a hypothesis $u_q : W \times Y \to [0,1]$.

d Calculate the pseudo-loss (for P and V_q):

$$\varepsilon_q = \sum_{(i,y): y_i \neq y} V_q(i)\left(1 - u_q(w_q, y_q) + u_q(w_q, y)\right).$$

e Calculate the weight update parameter:

$$\alpha_q = \frac{\varepsilon_q}{1 - \varepsilon_q}.$$

f Update V_q:

$$V_{q+1}(i) = V_q(i)\,\alpha_q^{\frac{1}{2}\left(1 + u_q(w_q, y_q) + u_q(w_q, y: y \neq y_q)\right)}.$$

g Normalize V_{q+1}: Let $Z_q = \sum_i V_{q+1}(i)$.

$$V_{q+1}(i) = \frac{V_{q+1}(i)}{Z_q}.$$

3 Output the final hypothesis:

$$U(w) = \underset{y \in Y}{\arg\max} \sum_{q=1}^{Q} u_q(w, y) \log \frac{1}{\alpha_q}.$$

Fig. 4 Algorithm of RUSBoost.

$$\begin{aligned} \mathrm{EI}_{\eta^*}(\xi) &= \int_{-\infty}^{+\infty} \max(\eta^* - \eta, 0) p_M(\eta \mid \xi) \mathrm{d}\eta \\ &= \frac{\int_{-\infty}^{\eta^*} (\eta^* - \eta) p(\eta) \mathrm{d}\eta}{\gamma + (1-\gamma) \dfrac{g(\xi)}{l(\xi)}} \end{aligned} \tag{9}$$

According to the formula, $\mathrm{EI}_{\eta^*}(\xi) \propto \left(\gamma + (1-\gamma)\dfrac{g(x)}{l(x)}\right)^{-1}$. As γ is a determined value, $\xi_{new} = \arg\max_{\xi} \dfrac{l(\xi)}{g(\xi)}$.

Step 3: The new observation point $D_{2:\,t+1} = \{\xi_{2:\,t+1}, \eta_{2:\,t+1}\}$ is added to the set of hyperparametric sample points.

Step 4: The above steps are repeated until a specified time bound is reached.

4 Results

Traffic events exhibit unique sound characteristics. The feature selection process for classifying traffic event sounds involved extracting multidimensional information from audio data based on the specific characteristics of each event type. The primary goals are to highlight the distinctiveness of different sound events and ensure comprehensive coverage of the feature space. SMBO-RUSBoost aimed to enhance both classification accuracy and interpretability.

The outcomes of the model require validation and interpretation to ensure accuracy and reliability. Although prediction accuracy serves as the most rudimentary metric for assessing model quality, it often fails to fully capture the efficacy of a model. To gauge the performance of the model impartially, we developed an elaborate, all-encompassing, and objective evaluation framework, as detailed in Table 2.

The dataset is split into 80% training and 20% test sets. The

Table 2 Calculation formula and meaning of the classification evaluation index

Index	Full name	Description	Formula
ACC(+)	Accuracy	Proportion of correctly classified samples to total	$\mathrm{ACC} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$
FDR(−)	False discovery rate	Proportion of samples among those classified as positive that are actually negative	$\mathrm{FDR} = \dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}$
TPR(+)	True positive rate	Proportion of actual positive samples that are correctly classified	$\mathrm{TPR} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
FPR(−)	False positive rate	Proportion of actual negative samples that are incorrectly classified	$\mathrm{FPR} = \dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$
PPV(+)	Positive predictive value	Proportion of samples among those classified as positive that are actually are positive	$\mathrm{PPV} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
FNR(−)	False negative rate	Proportion of actual positive samples that are incorrectly classified as negative	$\mathrm{FNR} = \dfrac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}$
TNR(+)	True negative rate	Proportion of actual negative samples that are correctly classified as negative	$\mathrm{TNR} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$
F1(+)	F1_score	The harmonic mean of PPV and TPR, which combines the accuracy and comprehensiveness of classification	$\mathrm{F1} = 2 \times \dfrac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}}$
ER(−)	Error rate	Proportion of incorrectly classified samples to total	$\mathrm{ER} = \dfrac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$
MCC(+)	Matthews correlation coefficient	Similarity and proximity between classified and actual samples. It has excellent objectivity for imbalance prediction	$\mathrm{MCC} = \dfrac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$
KAP(+)	Kappa	Metrics for consistency test and measuring classification accuracy	$\mathrm{KAP} = \dfrac{\mathrm{P_o} - \mathrm{P_e}}{1 - \mathrm{P_e}}$

Note: "+" represents "higher is better," "–" represents "lower is better". Implications of symbols: TP (True Positive): The model correctly predicted a positive outcome for a positive sample. TN (True Negative): The model correctly predicted a negative outcome for a negative sample. FP (False Positive): The model incorrectly predicted a positive outcome for a negative sample. FN (False Negative): The model incorrectly predicted a negative outcome for a positive sample. $\mathrm{P_o}$ (Predictive Consistency): This refers to the agreement between the observed outcomes and the predictions made by the model. $\mathrm{P_e}$ (Chance Consistency): This refers to the agreement that could be expected by chance alone.

hyperparameters of the SMBO-RUSBoost model are set as follows: the number of learners is 13, the learning rate is 0.7206, and the maximum number of splits is 679. To evaluate the model performance on the engineered dataset, several key metrics are used, including accuracy (ACC), error rate (ER), true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), false positive rate (FPR), F1 score (F1), Matthews correlation coefficient (MCC), and kappa statistic (KAP). Compared to other indicators, accuracy provides a macrolevel perspective, while TPR is the focus of our evaluation, as it evaluates the model according to the proportion of correct predictions within the positive class. The results of the study are shown in Table 3.

Table 3 Evaluation of the overall effectiveness of the model

	ACC	ER	TPR	TNR	PPV	FPR	F1	MCC	KAP
RUSBoost	0.689	0.311	0.520	0.894	0.512	0.106	0.495	0.455	0.028
SMBO-RUSBoost	0.965	0.035	0.918	0.989	0.973	0.011	0.940	0.933	0.891

To evaluate the performance improvement before and after model optimization, we compared the evaluation metrics of the RUSBoost and SMBO-RUSBoost models.

(1) The ACC of the SMBO-RUSBoost model increased from 0.689 to 0.965, while the ER decreased from 0.311 to 0.035, demonstrating an enhanced capability to distinguish different sound event categories.

(2) The TPR increased from 0.520 to 0.918, reflecting a significant improvement in correctly identifying positive class samples, such as specific traffic event sounds. Similarly, the TNR improved from 0.894 to 0.989, indicating better performance in correctly identifying negative class samples such as city noise.

(3) The PPV increased from 0.512 to 0.973, suggesting more accurate identification of positive samples and a reduced likelihood of misclassification. The FPR decreased from 0.106 to 0.011, highlighting a reduction in the misclassification of negative samples as positive.

(4) The F1 score improved from 0.495 to 0.940, indicating a better balance between the precision and recall. The MCC increases from 0.455 to 0.933, indicating a more balanced overall performance. The KAP increased from 0.028 to 0.891, demonstrating a strong alignment between the model predictions and actual results.

Further analysis of the classification accuracy for each traffic event before and after model optimization (see Table 4) revealed the following.

(1) In general, the SMBO-RUSBoost model exhibited excellent performance in classifying four special audio events, with both TPR and PPV exceeding 90.0%. This demonstrates the effectiveness of SMBO in hyperparameter optimization.

(2) After optimization, small-sample categories such as horn and siren showed substantial gains, rising from no positive predictions in the baseline model to PPV values near 99% in SMBO-RUSBoost. By contrast, noise and slide exhibited a slight decline in PPV, which may be attributed to the increased class sensitivity induced by model complexity, reflecting a typical trade-off between precision and recall.

(3) Following optimization, the TPR for crash events showed a marginal decrease, possibly because some crash samples were misclassified as noise by the SMBO-RUSBoost model. Moreover, owing to the complexity of noise, its TPR is lower than that of other events but still higher than 60%. Overall, the model performs reasonably well in classifying special road events and filtering common city noise.

The confusion matrix is shown in Fig. 5 from the figure. We observe that classification instability is a major problem in the RUSBoost model, likely due to the uneven distribution of the data. The imbalance, whether in the number of samples or the data contained within a single sample, contributes to the problem. Events with large amounts of training data, such as crashes and slides, tend to have higher classification accuracy, whereas events with smaller datasets, particularly horns and sirens, exhibit poor classification performance, with TPR of exactly 0.

Therefore, we compared the model performance before and after optimization using ROC curves. First, the FPR and TPR for each prediction model are calculated based on the confusion matrix. Next, we dynamically adjusted the thresholds, each corresponding to an FPR and TPR value. These points are connected to obtain the ROC curves (Fig. 6). Finally, the area under the curve (AUC) was determined (Table 4). The results demonstrate that the AUC for crashes, slides, and horns peaked before optimization, whereas that for siren and noise increased from 0.8 to 1.0, demonstrating the effectiveness of SMBO. It is worth noting that the predicted scores of positive and negative samples may differ in magnitude across classes or models, the ROC curves can exhibit different rising behaviors in the low false

Table 4 Classification accuracy for each traffic event

Classification of events	Norm	RUSBoost	SMBO-RUSBoost
Crash	AUC	1.00	1.00
	PPV (%)	57.1	93.2
	FDR (%)	42.9	6.8
	TPR (%)	99.9	99.6
	FNR (%)	0.1	0.4
Slide	AUC	1.00	1.00
	PPV (%)	100.0	99.7
	FDR (%)	0.0	0.3
	TPR (%)	99.4	99.6
	FNR (%)	0.6	0.4
Horn	AUC	1.00	1.00
	PPV (%)	—	98.9
	FDR (%)	—	1.1
	TPR (%)	0.0	97.8
	FNR (%)	100.0	2.2
Siren	AUC	0.88	1.00
	PPV (%)	—	99.8
	FDR (%)	—	0.2
	TPR (%)	0.0	96.9
	FNR (%)	100.0	3.1
Noise	AUC	0.80	0.99
	PPV (%)	99.0	94.9
	FDR (%)	1.0	5.1
	TPR (%)	60.7	64.9
	FNR (%)	39.3	35.1

Note: PPV and FDR for the horn and siren classes under the baseline model are reported as "—". This is because the classifier made no positive predictions for these minority classes, rendering both metrics mathematically undefined.

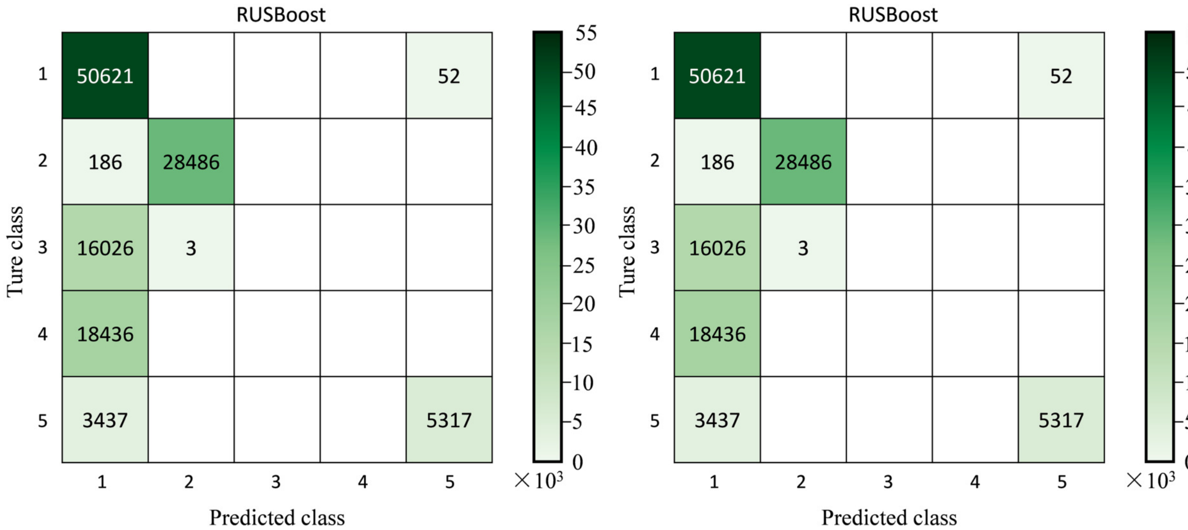


Fig. 5 Confusion matrix. The empty cells in the matrices correspond to class pairs with zero counts.

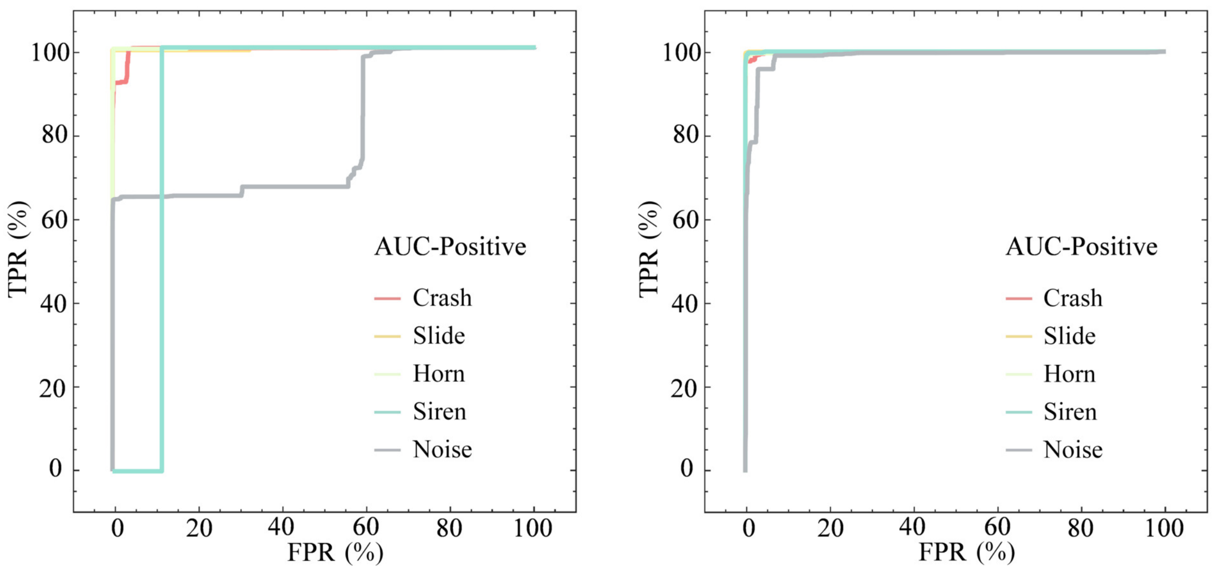


Fig. 6 ROC curve.

positive rate region, even when the overall ranking is perfect (yielding an AUC of 1). This implies that the local shape of the ROC curve may vary, while the total area under the curve (AUC) remains unchanged, resulting in differences in the upper-left portion of the ROC curve despite identical AUC values.

5 Conclusions

In conclusion, this study proposes and evaluates a sequence model-based optimization algorithm for classifying anomalous road events within the RUSBoost framework. Despite the presence of feature-similar events in imbalanced datasets, our method still achieves effective, accurate, and interpretable classification. The test results indicate that SMBO-RUSBoost achieves greater accuracy than traditional RUSBoost, highlighting the advantages of SMBO in parameter optimization.

A significant contribution of this study is the development of cost-effective and flexible classification methods for detecting abnormal road events, offering an alternative to existing road-monitoring techniques. Notably, the model achieved acceptable recognition accuracy in the study dataset, which included noisy data alongside abnormal events, indicating the resilience of the model to interference. This study aims to assist relevant authorities in swiftly identifying and responding to road accidents and safety risks.

Future research should focus on the application and testing of the models using more comprehensive datasets. Key areas for improvement include refining the sound classification and increasing the sample size. The proposed approach represents a significant methodological advancement and provides a more detailed and sensitive tool for road surveillance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 72261021), Science and Technology Innovation and Demonstration Projects, Department of Transport of Yunnan Province, China (Nos. 2023-121, 2022-23-3), Research Project on Key Technologies for Highway Safety and Disaster Mitigation and Prevention, Department of Transport of Yunnan Province, China (No. SZKM202231021), Science and Technology R&D Projects of Yunnan Communications Investment & Construction Group Co., Ltd. (No. YCIC- YF-2021-05). The experiments are carried out at the China Road Traffic Driving Simulation Laboratory, which is located at Kunming University of Science, China. All our participants should be thanked. In addition, we would like to thank the anonymous reviewers for their invaluable insights. Their thoughtful comments have been highly instrumental in improving the overall quality of this paper and providing valuable guidance for our future research.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Azimjonov, J., Özmen, A. 2021. A real-time vehicle detection and a novel vehicle tracking systems for estimating and monitoring traffic flow on highways. *Advanced Engineering Informatics*, **50**: 101393.

Boyarchikov, Y., Martinec, T. 2024. Road pavement monitoring and analysis approaches using gyroscope and accelerometer data. *IFAC-PapersOnLine*, **58**: 138–142.

Cui, Y., Wu, J., Xu, H., Wang, A. 2020. Lane change identification and prediction with roadside LiDAR data. *Optics & Laser Technology*, **123**: 105934.

Datondji, S. R. E., Dupuis, Y., Subirats, P., Vasseur, P. 2016. A survey of vision-based traffic monitoring of road intersections. *IEEE Transactions on Intelligent Transportation Systems*, **17**: 2681–2698.

Feng, Z., Ding, X., Jiang, Y. 2010. Pitch period estimation of voice signal based on EEMD and Hilbert Transform. In: Proceedings of the 2nd International Asia Conference on Informatics in Control, Automation and Robotics, 365–368.

Freund, Y., Schapire, R. E. 1996. Experiments with a new boosting algorithm. In: Proceedings of the International Conference on Machine Learning.

García-González, J., Molina-Cabello, M. A., Luque-Baena, R. M., Ortiz-de-Lazcano-Lobato, J. M., López-Rubio, E. 2021. Road pollution estimation from vehicle tracking in surveillance videos by deep convolutional neural networks. *Applied Soft Computing*, **113**: 107950.

Han, Z., Wang, J. 2015. Research on speech endpoint detection under low signal-to-noise ratios. In: Proceedings of the 27th Chinese Control and Decision Conference, 3635–3639.

Huang, X., Acero, A., Hon, H. W., Reddy, R. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR.

Ibili, F., Owolabi, A. O., Ackaah, W., Massaquoi, A. B. 2022. Statistical modelling for urban roads traffic noise levels. *Scientific African*, **15**: e01131.

Iglesias-Merchan, C., Laborda-Somolinos, R., González-Ávila, S., Elena-Rosselló, R. 2021. Spatio-temporal changes of road traffic noise pollution at ecoregional scale. *Environmental Pollution*, **286**: 117291.

Khan, S., Ali, H., Ullah, Z., Bulbul, M. F. 2019. An intelligent monitoring system of vehicles on highway traffic. In: Proceedings of the 12th International Conference on Open Source Systems and Technologies, 71–75.

Kim, B., Kim, D., Park, S., Jung, Y., Yi, K. 2016. Automated complex urban driving based on enhanced environment representation with GPS/map, radar, lidar and vision. *IFAC-PapersOnLine*, **49**: 190–195.

Kotov, A., Pospelov, P. 2017. Engineering tools and methods of estimation of traffic capacity using mobile video monitoring. *Transportation Research Procedia*, **20**: 347–354.

Lan, Z., Cai, M. 2021. Dynamic traffic noise maps based on noise monitoring and traffic speed data. *Transportation Research Part D: Transport and Environment*, **94**: 102796.

Liu, Y. Z., Tian, J. B. 2016. Double threshold endpoint detection algorithm based on speech enhancement. *Measurement & Control Technology*, **35**(11): 33–35. (in Chinese)

Liu, Y., Huang, W., Lin, X., Xu, R., Li, L., Ding, H. 2022. Variation of spatio-temporal distribution of on-road vehicle emissions based on real-time RFID data. *Journal of Environmental Sciences*, **116**: 151–162.

Marciniuk, K., Kostek, B. 2022. Machine learning applied to acoustic-based road traffic monitoring. *Procedia Computer Science*, **207**: 1087–1095.

Pascale, A., Guarnaccia, C., Macedo, E., Fernandes, P., Miranda, A. I., Sargento, S., Coelho, M. C. 2023. Road traffic noise monitoring in a Smart City: Sensor and Model-Based approach. *Transportation Research Part D: Transport and Environment*, **125**: 103979.

Piczak, K. J. 2015. ESC: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia, 1015–1018.

Pramanik, A., Sarkar, S., Maiti, J. 2021. A real-time video surveillance system for traffic pre-events detection. *Accident Analysis & Prevention*, **154**: 106019.

Prasad Das, C., Kumar Swain, B., Goswami, S., Das, M. 2021. Prediction of traffic noise induced annoyance: A two-staged SEM-Artificial Neural Network approach. *Transportation Research Part D: Transport and Environment*, **100**: 103055.

Por, E., Van Kooten, M., Sarkovic, V. 2019. Nyquist–Shannon sampling theorem. Leiden University.

Rabiner, L. R., Sambur, M. R. 1975. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, **54**: 297–315.

Rabiner, L. R. 1978. *Digital Processing of Speech Signals*. Pearson Education India.

Rahman, M. J., Zhu, H. 2023. Predicting accounting fraud using imbalanced ensemble learning classifiers–evidence from China. *Accounting & Finance*, **63**: 3455–3486.

Rathore, M. M., Paul, A., Rho, S., Khan, M., Vimal, S., Shah, S. A. 2021. Smart traffic control: Identifying driving-violations using fog devices with vehicular cameras in smart cities. *Sustainable Cities and Society*, **71**: 102986.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., Napolitano, A. 2010. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, **40**: 185–197.

Tedesco, S. A., Alexiadis, V., Loudon, W. R., Margiotta, R., Skinner, D. 1994. Development of a model to assess the safety impacts of implementing IVHS user services. In: Proceedings of the IVHS America Annual Meeting.

Thabit, A. S. M., Kerrache, C. A., Calafate, C. T. 2024. A survey on monitoring and management techniques for road traffic congestion in vehicular networks. *ICT Express*, **10**: 1186–1198.

Waggoner, B. 2009. Uncompressed video and audio: Sampling and quantization. In: *Compression for Great Video and Audio*. London: Routledge, 51–70.

Wang, X., Liu, H., Liu, Y. 2022. Auto-tuning deep forest for shear stiffness prediction of headed stud connectors. *Structures*, **43**: 1463–1477.

Wu, Z., Huang, N. E. 2009. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, **1**: 1–41.

Jiming Xie received his Ph.D. degree from Faculty of Transportation Engineering, Kunming University of Science and Technology, in 2024. As the first or corresponding author, he has published (including accepted) over 30 research papers, including more than 20 in SSCI/SCI/EI-indexed journals. He has been recognized as the top 5% Highly Cited Scholar by CNKI, with his work cited more than 500 times. His research interests include the coevolution of AI and human society, the ethical mechanisms of human-like intelligent agents, and sustainable complex systems. He has participated in more than 10 major research projects focused on system safety, including the National Key R&D Program of China, the National Natural Science Foundation of China, and provincial-level initiatives in social welfare and public services, technological innovation, and applied development

Yaqin Qin received her B.S. and M.S. degrees in mechanical engineering from Central South University, Changsha, China, and her Ph.D. degree in mechanical engineering from Kunming University of Science and Technology, Kunming, China, in 1995, 1998, and 2013, respectively. She had studied in UNBC in Canada and UT (Knoxville) in USA as visiting scholar. She is currently a professor with the College of Transportation Engineering, Kunming University of Science and Technology. she has authored more than 60 articles. Her research interests include transportation simulation, traffic human factors, intelligent transportation systems, and driving strategy in CAVs.

Yulan Xia is currently working as Ph.D candidate in University of Shanghai for Science and Technology. Her research interests include intelligent transportation systems, traffic flow theory, and traffic control and management.