

Acoustic Quality Evaluation Model and Optimization Design of Construction Machinery Products Based on Multi-Modal Perception

Jiayi Zhou *Swinburne College of Shandong University of Science and Technology, Jinan, Shandong, China****Corresponding author:** Jiayi Zhou. Email: 18514778188@163.com

Abstract

The subjective perception of noise quality in the cockpit of construction machinery is jointly determined by structural vibration and the radiated sound field, and the traditional evaluation method based on a single physical quantity finds it difficult to accurately describe its multi-dimensional coupling characteristics. This paper presents a sound quality evaluation model and optimization design method based on multi-modal sensing. Firstly, a multi-source acoustic and vibration synchronous acquisition test was designed, and psychoacoustic parameters, variational mode decomposition energy characteristics, and Gammatone cepstral coefficients were extracted to form acoustic features. The transmission path contribution rates and structural mode parameters were combined to form vibration features, and a multi-mode description system containing 102-dimensional original features was constructed. Secondly, a multi-head attention mechanism is introduced to achieve deep feature fusion, which is reduced to 38 dimensions after screening by the maximum-correlation minimum-redundancy (mRMR) criterion. On this basis, an improved Transformer evaluation model combining physical prior embedding of the transmission path, two-headed attention decoupling, and frequency-time-domain joint attention mechanism is proposed to predict the annoyance and preference scores. Furthermore, the feature contributions were quantified based on the SHAP method to locate the short board of acoustic quality, and the Gaussian process regression proxy model was constructed to replace the finite element simulation. The improved NSGA-III algorithm was used to achieve multi-objective optimization of plate thickness, suspension stiffness, and reinforcement layout. Taking loaders, rollers, and excavators as validation objects, the results show that the mean absolute error of the evaluation model on 1,440 independent test samples is 0.31-0.40, which is about 50% lower than that of the traditional psychoacoustic model. After optimization, the comprehensive improvement index reaches 0.52-0.91, and the prediction deviation of the proxy model is controlled within 0.30. This study provides data-driven methodology support for forward design and rapid evaluation of sound quality of construction machinery.

Keywords

Sound quality evaluation; Multi-mode sensing; Transformer network; Transfer path analysis; Multiple objective optimization

Received: 21 Jun 2026; Revised: 18 Jul 2026; Accepted: 29 Jul 2026; Published: 25 Aug 2026

Copyright: ©2026 The Author(s). Published by [Digital Intelligence Press Limited](#). This is an open access article under the [CC BY 4.0](#) license.

1. INTRODUCTION

Construction machinery plays an irreplaceable key role in infrastructure construction, mining and municipal engineering. The acoustic environment inside the cab of the construction machinery directly affects the work efficiency and occupational health of the operators. Traditional noise control has long

used A-weighted sound pressure level as the core indicator, but this indicator cannot effectively characterize subjective auditory perception attributes such as annoyance and comfort [1]. This has prompted the research on sound quality to shift from the “noise control” paradigm to the “quality design” paradigm. The sound quality problem of construction machinery is far more complex than that of passenger cars. There are multiple sources of excitation such as engines, transmissions and hydraulic pumps. The propagation of structural paths and airborne sound paths are intertwined. Moreover, the transient changes in working conditions make the sound field exhibit strong non-stationarity. It is urgent to establish a special analysis framework for multi-source and multi-path scenarios.

Existing sound quality evaluation research has formed three technical routes: psychoacoustic parameter quantification, machine learning prediction modeling and subjective evaluation verification. Psychoacoustic parameters (loudness, sharpness, etc.) are highly simplified approximations of the human hearing mechanism [2]. They have a complete description defect for the noise of construction machinery with impact and modulation characteristics [3],[4]. Machine learning methods, utilizing support vector machines, random forests, and various neural networks, learn subjective rating mapping from time-frequency features, which to some extent compensate for information loss [5],[6],[7]. However, most studies are still limited to single sound signals and do not fully consider structural vibration as a supplementary information source of noise radiation. In fact, the sound pressure level near the driver's ear in the cab is the result of the combined effects of structural vibration through multiple paths such as the suspension system, frame, and panels. If only the sound pressure level is analyzed without tracing the vibration source, designers cannot accurately locate the structural causes when sound quality problems occur.

The introduction of the concept of multimodal perception provides a new perspective for overcoming the above dilemmas. Multimodal perception emphasizes using heterogeneous sensors to obtain representational information of the same physical process in different domains and achieving a complete scene understanding that is difficult to achieve under single-modal conditions through information fusion [8],[9],[10],[11]. In sound quality problems, sound pressure and structural vibration are physical mappings of the sound-vibration coupling process in the fluid and solid domains, possessing definite causality and frequency domain coherence [12]. However, a methodology for systematically integrating vibration information into the evaluation model and guiding structural optimization has not yet been formed in sound quality evaluation [13],[14],[15]. Existing optimization studies mostly adopt a two-stage separation model of "evaluation first, optimization later", lacking a unified information transmission link between evaluation and optimization.

To address this gap, this paper, based on the physical coupling mechanism of acoustic vibration, integrates the concept of multimodal perception throughout the entire chain of feature expression, evaluation modeling, and optimization decision-making. First, a framework for synchronous acoustic vibration acquisition and deep fusion of multimodal features is established, so that the model input simultaneously includes acoustic properties and physical information of structural transmission paths. Second, an evaluation model is constructed with an improved Transformer network as the core, and the generalization ability under small sample conditions is enhanced by physical prior embedding and functional decoupling of attention mechanism [16]. Finally, the results of sensitive parameter analysis are fed back to the multi-objective structural optimization based on the surrogate model, upgrading the acoustic quality evaluation from “discovering problems” to “solving problems”. The research route of this paper reflects a complete closed loop from physical perception to cognitive evaluation and then to decision optimization, aiming to provide a quantifiable and traceable technical path for the positive design of acoustic quality in engineering machinery.

2. MULTI-MODE PERCEPTION FEATURE EXTRACTION AND FUSION METHOD ORIENTED TO SOUND QUALITY

2.1 Experimental design of multi-source acoustic vibration signal acquisition for construction machinery

The test object selects a certain type of wheel loader as a representative of typical construction machinery, which has significant characteristics of complex structure, changeable working conditions and harsh acoustic environment of the cab, which can fully reflect the typicality and challenge of the acoustic quality problem of construction machinery. At the same time, a hydraulic excavator of the same tonnage level is selected as the auxiliary verification object to investigate the generalization and migration ability of the proposed method among different models. The working condition design covers the four most representative operating states in the actual operation of construction machinery: Idling condition (the rated speed of the engine is about 650-750r/min, No load action), constant speed driving condition (gearbox second gear, speed about 8-10 km/h), full load acceleration condition (bucket full load, engine speed from idle to rated speed), and typical V-type working cycle condition (including shoveling, lifting, heavy load transportation, unloading, no load return five continuous working sections). The data collection time under each working condition is no less than 120 seconds, and the working cycle condition is recorded for no less than 5 complete cycles to ensure that the sample size is sufficient to support the training requirements of the subsequent deep learning model.

Sensor layout is the key link to determine the completeness of multi-mode information. The acoustic signal was collected by using a paraear microphone array composed of six 1/2 inch free-field microphones, four of which were arranged at about 150 mm beside the driver's left and right ears (two on the left and right sides were backup for each other), one was arranged at the center of the top of the cab to monitor the sound field distribution inside the car, and the other was arranged at about 1 m outside the engine compartment to monitor the near-field radiated noise. Twelve three-dimensional piezoelectric acceleration sensors are used for vibration signal acquisition, which are respectively arranged at the cab suspension mounting point (one for each of the four mounting points), the engine front suspension and rear suspension body side (one for each), the surface of the gearbox housing, the front axle and rear axle housing, and the center position of the key plates of the cab floor and side circumference. The speed pulse signal is picked up by the magnetoelectric speed sensor to pick up the signal of the engine flywheel gear ring, and the real-time speed data of the engine ECU is read by the CAN bus as redundancy verification. All sensor channels were sampled synchronously by the same set of multi-channel data acquisition system, and the sampling frequency was uniformly set as 51.2 kHz to meet the Nyquist sampling requirements of the upper limit frequency of construction machinery noise of 20 kHz. The cut-off frequency of anti-aliasing filter in vibration channel was set as 20 kHz, and the bandpass filter in acoustic channel was 20 Hz-20 kHz. The synchronous triggering mode is hardware triggering, and the time delay between channels is strictly controlled within $\pm 1\mu\text{s}$.

The test environment and data acquisition specifications strictly comply with GB/T 3785.1-2010 and ISO 6394:2008 and other relevant standards. The experiment was conducted in A semi-anechoic chamber with background noise below 20 dB (A) and ambient temperature controlled at $20\pm 2^\circ\text{C}$. Before the formal collection of the test sample, run the heat engine at the rated working condition for at least 30 minutes, so that the oil temperature, water temperature and friction pair temperature reach a stable working state. Wait 5 seconds for the system to stabilize after each working condition switch before starting acquisition. To ensure the repeatability and statistical robustness of the data, the collection was

repeated three times for each working condition, and each time the interval was not less than 10 minutes for system cooling recovery. During the acquisition process, the signal quality of each channel is monitored in real time, and abnormal data segments caused by cable loosening or sensor shedding are eliminated. The signal-to-noise ratio of the final effective data is required to be no less than 60 dB for acoustic channel and no less than 40 dB for vibration channel.

2.2 Construction of multi-dimensional representation system of acoustic features

Extracting acoustic features with auditory relevance and physical interpretability from the acquired parotid sound pressure signals is the first step to establishing the mapping relationship between objective physical quantities and subjective perceptions. In this paper, three hierarchical acoustic feature systems are constructed: the traditional psychoacoustic parameters provide the top-level measurement in line with the human auditory perception mechanism, the time-frequency domain fine features reveal the local non-stationary structure and order components of the signal, and the auditory spectral features simulate the frequency selection characteristics of the basilar membrane to retain the complete spectral envelope information.

The traditional psychoacoustic parameters are calculated according to international standards such as Zwicker loudness model and Fastl sharpness model. The loudness of the steady state signal was calculated according to ISO 532-1 (Zwicker method), and the time-varying loudness was calculated according to ISO 532-2 (Mool-Glasberg method) for the non-steady state condition. Specifically, the characteristic frequency band loudness $N'(z)$ is obtained by applying equal-response curve correction and band width normalization to the 1/3 octave sound pressure level $L(i)$:

$$N' = \sum_{i=1}^{24} \left[0.08 \times 10^{0.1L(i) + 0.05 \sum_{j=1}^{i-1} \Delta L(j)} \right]^{0.23} \quad (1)$$

Where, i and j are frequency band serial numbers on the Bark scale, ranging from 1 to 24, corresponding to the division of human auditory frequency bands; $L(i)$ is the sound pressure level (in dB) in the i -th Bark frequency band; $\Delta L(j)$ is the correction term of sound pressure difference between adjacent frequency bands; 0.23 is the loudness conversion index, derived from Stevens power law. The total loudness N is integrated over the full frequency band N' , and its unit is sone. The sharpness S reflects the degree of auditory "harshness" caused by the proportion of high frequency energy. The frequency band number $g(z)$ is integrated with the weight of the specific loudness $N'(z)$:

$$S = 0.11 \times \frac{\int_0^{24\text{Bark}} N'(z) \cdot g(z) \cdot z \cdot dz}{\int_0^{24\text{Bark}} N'(z) \cdot dz} \quad (2)$$

In the formula, z is the bark-scale frequency band serial number (dimensionless, numerical range 0-24), $g(z)$ is the frequency band weighting function, which takes the value of 1 in the frequency band above 15 Bark and the exponential decay below 15 Bark; 0.11 is the dimensional conversion factor to normalize the sharpness unit to acum. Roughness R reflects the modulation amplitude of the signal in the range of 70-300 Hz, and its calculation is based on modulation frequency f_{mod} and modulation depth ΔL_{mod} :

$$R = 0.3 \cdot f_{mod} \cdot \int_0^{24\text{Bark}} \Delta L_{mod}(z) \cdot dz \quad (3)$$

Where $\Delta L_{mod}(z)$ is the modulation depth (in dB) in the z -th Bark band, f_{mod} is the modulation frequency (in Hz), and 0.3 is the empirical constant (in $\text{asper} \cdot \text{Hz}^{-1} \cdot \text{dB}^{-1}$). The fluctuation intensity F is calculated for a lower modulation frequency range (about 0.5-20 Hz), using the following relationship:

$$F = \frac{1}{4} \cdot \frac{\Delta L_{total}}{f_{mod}} \cdot \left[1 + \sqrt{\frac{2\pi f_{mod}}{0.02}} \right]^{-1} \quad (4)$$

Where ΔL_{total} is the total modulation depth (in dB) and 0.02 is the empirical value of the auditory time constant (in s). The physical significance of this equation lies in the nonlinear change of the perception threshold of the fluctuation intensity with the modulation frequency.

Variational mode decomposition (VMD) and order tracking are combined for fine feature extraction in time-frequency domain. VMD adaptively decomposes the original sound pressure signal $p(t)$ into K eigenmode functions $u_k(t)$, each mode closely around its central frequency ω_k . VMD achieves decomposition by solving the following constrained variational problem:

$$\min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \quad (5)$$

The constraint condition is $\sum_{k=1}^K u_k(t) = p(t)$. Where $u_k(t)$ is the k -th modal component, ω_k is its corresponding center frequency (unit rad/s), $\delta(t)$ is the Dirac delta function, $*$ is the convolution operator, j is the imaginary unit, and ∂_t is the first-order partial derivative with respect to time. The solution of the variational problem is transformed into an augmented Lagrange function by introducing a quadratic penalty factor α and a Lagrange multiplier $\lambda(t)$, which is iteratively updated by the alternating direction multiplier method. The normalized energy ratio E_k of each modal component is calculated as follows:

$$E_k = \frac{\int |u_k(t)|^2 dt}{\sum_{m=1}^K \int |u_m(t)|^2 dt} \quad (6)$$

Where m is the modal index variable with values ranging from 1 to K . In order tracking, the speed pulse signal is used to resample the time domain signal into the angular domain stationary signal, so as to extract the order spectrum synchronized with the speed of the rotating component. The order O is defined as the ratio of frequency to the speed frequency of the reference shaft:

$$O = \frac{60 \cdot f}{n} \quad (7)$$

Where f is the frequency (in Hz) and n is the speed of the reference shaft (in r/min). The order spectrum is obtained by fast Fourier transform of the resampled angular domain signal, and the amplitude A_O of the engine ignition order (corresponding to the 2nd, 4th, 6th and 8th orders of the four-cylinder and four-stroke engine) and the gearbox engagement order are selected as the key order features.

Auditory spectral feature extraction aims to simulate the frequency resolution characteristics of

human ear basilar membrane. Mel spectrum maps linear frequency to Mel frequency scale to conform to the nonlinear perception of pitch by human ear, and its conversion relationship is as follows:

$$f_{mel} = 2595 \cdot \log_{10} \left(1 + \frac{f}{700} \right) \quad (8)$$

Where f_{mel} is the Mel frequency (in Mel) and f is the linear frequency (in Hz). The power spectrum $P(f)$ is obtained by short-time Fourier transform after frame and window of sound pressure signal $p(t)$, and the energy output M_b of each Mel frequency band is obtained by weighted sum of Mel triangular filter bank:

$$M_b = \sum_{f=0}^{F_s/2} P(f) \cdot H_b(f), \quad b = 1, 2, \dots, B \quad (9)$$

$H_b(f)$ is the frequency response function of the b -th Mel filter, B is the number of filter banks ($B = 40$ in this paper), and F_s is the sampling frequency (unit Hz). The dynamic characteristics of Mel spectrum are further logarithmically taken, and then discrete cosine transform is performed to obtain Mel frequency cepstrum coefficient (MFCC) [17],[18]. Different from MFCC, Gammatone cepstral coefficient adopts Gammatone filter bank that is closer to the impulse response of human ear basilar membrane, and its time domain impulse response $g(t)$ is:

$$g(t) = a \cdot t^{m-1} \cdot e^{-2\pi b t} \cdot \cos(2\pi f_c t + \varphi) \quad (10)$$

Where a is the gain normalization constant, t is the time (unit s), m is the filter order (often taken as 4), b is the attenuation factor (unit Hz), f_c is the center frequency (unit Hz), and φ is the initial phase. The calculation process of GFCC is similar to that of MFCC, but it has better robustness to the common impact components in construction machinery noise in terms of spectral envelope representation [19],[20]. For each acoustic analysis frame (frame length 1024 points, frame shift 512 points), 12-dimensional MFCC, 12-dimensional GFCC, 4-dimensional psychoacoustic parameters, energy and order amplitude of each modal component were extracted. The final acoustic feature dimension of a single sample was $D_a = 48$.

2.3 Feature extraction of vibration-structure transfer path

As the physical source of noise radiation, the transmission path characteristics of structural vibration directly determine the spectral structure of radiated noise. In this paper, the transfer path analysis (TPA) method is introduced to quantify the contribution of each excitation source to the auricular sound field of the cab through the structural path, and the dominant transmission channel of vibration energy is revealed through the structural mode analysis [21],[22]. Based on the superposition principle of linear time-invariant system, the total sound pressure $P_{target}(f)$ of the target point around the driver's ear is expressed as the linear superposition of the contribution of each excitation source $F_i(f)$ through the transfer function $H_{TP,i}(f)$:

$$P_{target}(f) = \sum_{i=1}^{N_s} H_{TP,i}(f) \cdot F_i(f) \quad (11)$$

Where N_s is the number of excitation sources, including engine excitation, transmission excitation, axle housing excitation and hydraulic pump excitation and other main vibration sources;

$F_i(f)$ is the force of the i -th excitation source at the connection point (unit N); $H_{TP,i}(f)$ is the transfer function (unit Pa/N) from the i -th excitation source to the target sound field, which is measured by hammering method or shaker method in the shutdown state. In this paper, the OSPA method is adopted, and the vibration acceleration $a_j(f)$ collected under actual operating conditions is used as a reference signal to replace the excitation force $F_i(f)$, which is difficult to measure directly, and the following regression equation is established [23]:

$$P_{target}(f) = \sum_{j=1}^{N_r} H_{OTPA,j}(f) \cdot a_j(f) \quad (12)$$

N_r is the number of reference channels (12 vibration measurement points in this paper), $a_j(f)$ is the frequency domain acceleration signal (unit m/s²) of the j -th acceleration measurement point, and $H_{OTPA,j}(f)$ is the transmission path coefficient of working conditions. The contribution $C_j(f)$ of each path is defined as:

$$C_j(f) = H_{OTPA,j}(f) \cdot a_j(f) \quad (13)$$

The frequency band energy proportion η_j contributed by each path is further defined as follows:

$$\eta_j = \frac{\int_{f_1}^{f_2} |C_j(f)|^2 df}{\sum_{k=1}^{N_r} \int_{f_1}^{f_2} |C_k(f)|^2 df} \times 100\% \quad (14)$$

f_1 and f_2 are the lower and upper limits of the frequency of concern respectively (20 Hz and 2000 Hz are taken in this paper), and k is the path index variable. The multi-reference point least squares complex frequency domain method (PolyMAX) was used to identify the structural modal parameters. The frequency response function matrix $H_{pq}(\omega)$ of the steady-state vibration response signal $s(t)$ can be expressed as a rational fraction form in the Laplace domain s (unit s⁻¹):

$$H_{pq}(s) = \sum_{r=1}^{N_m} \left(\frac{A_{pqr}}{s - \lambda_r} + \frac{A_{pqr}^*}{s - \lambda_r^*} \right) + \frac{e_{pq}}{s^2} + \frac{d_{pq}}{s} \quad (15)$$

Where p and q are the indexes of response points and excitation points respectively, N_m is the order of the identified mode, A_{pqr} is the residue matrix element of the r -th mode, λ_r is the complex extreme value, $\lambda_r = -\xi_r \omega_{nr} + j\omega_{nr}\sqrt{1 - \xi_r^2}$, ξ_r is the modal damping ratio (dimensionless), ω_{nr} is the undamped natural frequency in rad/s, $*$ denotes the complex conjugate, and e_{pq} and d_{pq} are the residual compliance and residual mass terms, respectively. The mode shape participation factor ϕ_{jr} of each mode is obtained from the residue matrix decomposition:

$$A_{pqr} = \phi_{pr} \cdot \Gamma_{qr} \quad (16)$$

ϕ_{pr} is the mode shape component of the r -th mode at the response point p , and Γ_{qr} is the mode participation coefficient of the r -th mode at the excitation point q . The frequency ω_{nr} , damping ratio ξ_r and mode shape amplitude ϕ_{jr} of the first $N_m = 15$ dominant modes are extracted from the identified modal parameters as structural dynamics characteristics.

The coupling relationship between structural vibration characteristics and radiated noise is

quantitatively described by the coherence function and the cross-power spectral density. The amplitude squared coherence function $\gamma_{pa}^2(f)$ between sound pressure $p(t)$ and vibration acceleration $a(t)$ is defined as follows:

$$\gamma_{pa}^2(f) = \frac{|G_{pa}(f)|^2}{G_{pp}(f) \cdot G_{aa}(f)} \quad (17)$$

Where $G_{pa}(f)$ is the mutual power spectral density of sound pressure and vibration acceleration (unit $\text{Pa} \cdot \text{m}^2/\text{s}^3$), $G_{pp}(f)$ is the self-power spectral density of sound pressure (unit Pa^2/Hz), $G_{aa}(f)$ is the self-power spectral density of vibration acceleration (unit $\text{m}^2/\text{s}^4/\text{Hz}$). The value of $\gamma_{pa}^2(f)$ is between 0 and 1, and the closer it is to 1, the more linearly determined the sound pressure in this frequency band is by the signal of the vibration measurement point. Further define the coupling coefficient κ_j of each vibration measurement point to the parotid sound pressure as the weighted average of the coherence function in the concerned frequency band:

$$\kappa_j = \frac{\int_{f_1}^{f_2} W(f) \cdot \gamma_{pa,j}^2(f) \cdot df}{\int_{f_1}^{f_2} W(f) \cdot df} \quad (18)$$

$W(f)$ is the frequency weighting function (dimensionless) of the A-weight network, which is used to emphasize the 1-6 kHz frequency band that the human ear is most sensitive to. By combining the κ_j of each measurement point with the TPA contribution rate η_j , the advantage transfer path with significant contribution and coupling degree is screened out. For each sample, the contribution rate of each path, the frequency and damping ratio of the first 15 modes, and the coupling coefficient of each measurement point are extracted to form the vibration structure feature vector, and the dimension is $D_v = N_r + 2N_m + N_r = 54$ (12 dimensions of path contribution rate, 15 dimensions of frequency, 15 dimensions of damping ratio, and 12 dimensions of coupling coefficient).

2.4 Multi-modal feature alignment and deep fusion strategy

There are significant differences between acoustic features and vibration structure features in sampling rate, time scale and physical dimension, so direct splicing or simple weighted fusion will lead to low efficiency of information utilization. In this section, the time synchronization and dimension alignment of the two types of features are firstly carried out, and then the multi-mode deep fusion architecture based on multi-head attention mechanism is designed. Finally, feature selection and dimension reduction are realized through information redundancy analysis.

The core of time synchronization and dimension alignment is to establish time correlation between frame level acoustic features and corresponding vibration structure features. Since vibration transfer path analysis is usually carried out in the frequency domain, while acoustic psychoacoustic parameters are calculated in the time domain, this paper uses equally spaced frames to divide and unify the time basis of the two features. Set the original sampling frequency as $F_s = 51200$ Hz, frame length $L = 2048$ points, frame shift $\Delta L = 1024$ points, then the frame period T_{frame} is:

$$T_{frame} = \frac{\Delta L}{F_s} = \frac{1024}{51200} = 0.02 \text{ s} \quad (19)$$

That is, a feature vector is output every 20 ms. For the same time frame $t_n = n \cdot T_{frame}$ (n is the frame number), the acoustic feature vector $a(t_n) \in \mathbb{R}^{48}$ and the vibration structure feature vector

$v(t_n) \in \mathbb{R}^{54}$ are calculated simultaneously. In order to achieve the consistency of feature dimensions for subsequent attention calculation, two groups of trainable linear projection matrices $W_a \in \mathbb{R}^{d_m \times 48}$ and $W_v \in \mathbb{R}^{d_m \times 54}$ are introduced, where d_m is the common embedding dimension after projection ($d_m = 64$), and the two types of features are projected to the same feature space:

$$X_a = W_a \cdot a, X_v = W_v \cdot v \quad (20)$$

a and v are the acoustic and vibration eigenvectors (column vectors), respectively, and X_a and X_v are the embedding vectors ($\in \mathbb{R}^{64}$) after projection. At the same time, in order to retain the dependence of features on the time axis, the position coding of the projected feature sequence is carried out:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_m}}}\right), PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_m}}}\right) \quad (21)$$

pos is the position index of the frame in the time series, and i is the dimension index of the embedding vector (the value ranges from 0 to 31). After encoding, the feature of each position carries its temporal context information.

The core idea of multi-head attention mechanism is to allow the model to simultaneously capture multiple dependencies between different modes and different time steps in different subspaces. Let $d_k = d_m/h$ be the dimension of each attention head, where h is the number of attention heads ($h = 8$ in this paper), then the attention output of the r -th head is:

$$\text{Attention}_r(Q_r, K_r, V_r) = \text{softmax}\left(\frac{Q_r K_r^T}{\sqrt{d_k}}\right) V_r \quad (22)$$

Q_r (query matrix), K_r (key matrix) and V_r (value matrix) are obtained by linear transformation of X_a and X_v :

$$Q_r = XW_{Qr}, K_r = XW_{Kr}, V_r = XW_{Vr} \quad (23)$$

$X = [X_a, X_v] \in \mathbb{R}^{d_m \times 2}$ is the modal feature matrix after splicing (the number of columns is 2), and $W_{Qr}, W_{Kr}, W_{Vr} \in \mathbb{R}^{d_m \times d_k}$ is the learnable weight matrix of the r -th attention head. $\sqrt{d_k}$ in softmax function is the scaling factor, which is used to avoid gradient saturation caused by too large dot product. The final output O of multi-head attention is the projection of the output of each head after splicing:

$$O = [\text{Attention}_1, \text{Attention}_2, \dots, \text{Attention}_h] W_O \quad (24)$$

$W_O \in \mathbb{R}^{d_m \times d_m}$ is the output projection matrix. On this basis, this paper introduces a cross-attention enhancement mechanism between modes, so that acoustic features are used as query modes to query vibration features, and vibration features are used as query modes to query acoustic features. Two-way information flow between modes is calculated through cross-attention:

$$O_{a \rightarrow v} = \text{Attention}(Q_a, K_v, V_v), O_{v \rightarrow a} = \text{Attention}(Q_v, K_a, V_a) \quad (25)$$

Q_a comes from the linear projection of X_a , K_v and V_v come from the linear projection of X_v , and vice versa. The final fusion feature F_{fusion} adaptively combines the output of bidirectional cross-attention through the gating mechanism:

$$F_{fusion} = g \odot O_{a \rightarrow v} + (1 - g) \odot O_{v \rightarrow a} \quad (26)$$

$g \in [0,1]^{d_m}$ as the gating vector (by learning the parameters of the $W_g \in \mathbb{R}^{d_m \times 2d_m}$ $g = \sigma(W_g[O_{a \rightarrow v}; O_{v \rightarrow a}])$, the σ as Sigmoid activation function, even $\odot$ for product line by element, $[\cdot; \cdot]$ said vector splicing. The gating mechanism enables the model to adaptively adjust the contribution weight of different modal features according to the difference in the amount of modal information currently input.

Feature redundancy analysis and selection aim to eliminate highly correlated components in multimodal features and reduce feature dimension to reduce computational cost while maintaining information quantity. The maximum correlation minimum redundancy (mRMR) criterion based on mutual information is used for feature selection, and the mutual information $I(f_i; y)$ between feature f_i and target subjective score y is calculated as follows [24],[25],[26]:

$$I(f_i; y) = \sum_{x_i \in X_i} \sum_{y \in Y} P(x_i, y) \log \frac{P(x_i, y)}{P(x_i)P(y)} \quad (27)$$

$P(x_i, y)$ is the joint probability distribution of the feature value x_i and the target score y , and $P(x_i)$ and $P(y)$ are their respective marginal probability distributions. The redundancy between two features f_i and f_j is measured by mutual information $I(f_i; f_j)$. The screening criteria of mRMR maximize correlation while minimizing redundancy:

$$\max_{\mathcal{S}} \left[\frac{1}{|\mathcal{S}|} \sum_{f_i \in \mathcal{S}} I(f_i; y) - \frac{1}{|\mathcal{S}|^2} \sum_{f_i, f_j \in \mathcal{S}} I(f_i; f_j) \right] \quad (28)$$

$\mathcal{S}$ is the selected subset of features, $|\mathcal{S}|$ is the subset size. The forward sequential selection strategy is adopted for all $D_{total} = D_a + D_v = 102$ dimensional features, and the features that maximize the objective function are added one by one starting from the empty set, which is terminated when the addition of new features no longer causes the objective function to grow beyond the threshold $\epsilon = 10^{-3}$. The number of features retained after mRMR screening is about 38 dimensions, and the compression rate is 62.7%. At the same time, the average redundancy $\bar{R}_{intra}$ of the filtered feature set is calculated as the evaluation index:

$$\bar{R}_{intra} = \frac{2}{N_s(N_s - 1)} \sum_{i=1}^{N_s} \sum_{j=i+1}^{N_s} \rho_{ij} \quad (29)$$

N_s is the number of features after screening, and ρ_{ij} is the Pearson correlation coefficient between the i -th and j -th features. Finally, the selected fusion feature vector $F_{selected} \in \mathbb{R}^{38}$ is used as the input of the sound quality evaluation model in section 3, which comprehensively represents the whole picture of the sound quality of construction machinery in the way of physical diversity, perceptual complementarity and low redundancy at the information level.

Table 1. Summary of acoustic feature extraction methods and feature dimensions at all levels

Level of feature	Method of extraction	Core calculation basis	Dimension of feature
Psychoacoustic parameters	Zwicker loudness model, sharpness model	Bark band sound pressure level integration, specific loudness weighting	4
Time-frequency fine characteristic	VMD decomposition, order tracking	Variational mode decomposition energy ratio, order amplitude	8
Auditory spectrum feature	Mel filter bank, Gammatone filter	Nonlinear frequency mapping, basilar membrane impulse response	24
Transmission path characteristics	OTPA, PolyMAX modal analysis	Path contribution rate calculation and rational fraction fitting of frequency response function	42

[Table 1](#) summarizes the extraction methods and corresponding dimensions of acoustic features at each level, among which the dimension of transmission path feature is the highest (42 dimensions), reflecting the importance of structural transmission information to the representation of acoustic quality.

Table 2. Key parameter configuration of multi-modal feature fusion architecture

Name of parameter	Symbols	Numerical value	Basis of setup
Common embedding dimension	d_m	64	The geometric meaning of the original dimensions of the two types of features is approximated
Head of attention	h	8	Make $d_k = 8$ for each head dimension to balance calculation and representation ability
Single-head key value dimension	d_k	8	$d_k = d_m/h$, Keep the total number of parameters moderate
Fusion feature dimension	D_{fusion}	64	Consistent with d_m , it is easy to connect the residuals
mRMR screening threshold	ϵ	10^{-3}	The selection is terminated when the mutual information increment is less than this value
Feature dimension after screening	N_s	38	Automatically determined by the mRMR criterion

[Table 2](#) lists the key parameter configuration of multimodal fusion architecture. All parameters are set based on the pre-experimental results of the previous ablation experiment to ensure a reasonable balance between model expression ability and computational efficiency. The input of the above feature extraction and fusion process is the multi-source signal of sound pressure-vibration-speed collected synchronously, and the output is the low-dimensional fusion feature vector $F_{selected}$ after time alignment, dimension projection, attention fusion and redundancy screening. The whole process realizes a triple leap from physical signal to perceptual feature, from single-mode description to multi-mode complementarity, and from high-dimensional redundancy to low-dimensional compactness.

3. SOUND QUALITY EVALUATION MODEL BASED ON IMPROVED DEEP NETWORK

3.1 Construction of subjective evaluation database of sound quality

Subjective evaluation databases are the core support of supervised learning. In this paper, a multi-modal feature database with subjective labeling is established for the sound quality of construction machinery cabs by using the mixed test strategy of grade scoring method and pair comparison method. The evaluation object is the acoustic vibration synchronization signals collected under four typical working conditions in section 2. For each working condition, 20 representative 6-second sample segments are selected, totaling 80 samples. The evaluation dimension is set as subjective annoyance (t) and preference (t). The former reflects the operator's dissatisfaction with the current acoustic environment, and the latter reflects the relative liking compared with other samples. Both of them adopt 7-point Likert scale (1 is excellent/no annoyance at all; 7 rated as extremely poor/extremely annoyed).

The selection and training of evaluators are strictly in accordance with ITU-R BS.1116-3 standards. A total of 35 candidate evaluation personnel were selected from the R&D center, after-sales service department and operation training school of a construction machinery manufacturing enterprise. All candidates were required to pass the hearing threshold test (the hearing threshold of 500, 1000, 2000 and 4000 Hz was not higher than 20 dB HL) and tone discrimination test. The training stage is divided into two parts: the first part is anchoring point sample calibration. Three pre-calibrated anchor point samples of excellent (sound quality score close to 1), poor (sound quality score close to 7) and medium (sound quality score about 4) are played to the evaluators, so as to establish a unified scoring reference system. The second part was repetitive training, in which 10 samples were randomly selected and played twice in a random order. The evaluators whose absolute value of the difference between the two scores exceeded 1 were eliminated. After screening, 24 evaluators were finally retained, including 14 with experience in construction machinery operation, 6 acoustic engineers and 4 general listeners. The gender ratio was about 2:1, and the age distribution was 25-48 years old.

The formal evaluation test was conducted in A listening room with background noise less than 25 dB (A) and reverberant time less than 0.3 seconds. A total of $C(80,2) = 3160$ pairs of combinations were performed on all 80 samples by using the pairwise comparison method, and each pair of samples was played twice (AB and BA) in random order. In order to avoid auditory fatigue, the whole experiment was divided into 8 sessions, with an interval of at least 2 hours for each session and a single session lasting no more than 35 minutes. The subjective annoyance score $S_{annoy}^{(i)}$ is obtained by using Thurstone-Mosteller rule to transform the pairwise comparison results, while the preference score $S_{pref}^{(i)}$ is directly obtained from the rating score results. The intra-group correlation coefficient is used to calculate the two-way random ICC (2,1) value of the scores of 24 evaluators in the reliability consistency test. The final number of reliable samples was 73 (excluding 7 samples with insufficient reliability), and the final subjective score of each sample was calculated as the arithmetic mean of the scores of 24 evaluators $\bar{S}^{(i)}$:

$$\bar{S}^{(i)} = \frac{1}{N_p} \sum_{m=1}^{N_p} S_m^{(i)} \quad (30)$$

Where $N_p = 24$ is the number of effective evaluators, and $S_m^{(i)}$ is the score of the m -th evaluator on the i th sample. At the same time, the standard deviation $\sigma_S^{(i)}$ of each sample score is calculated:

$$\sigma_S^{(i)} = \sqrt{\frac{1}{N_p - 1} \sum_{m=1}^{N_p} \left(S_m^{(i)} - \bar{S}^{(i)} \right)^2} \quad (31)$$

As a measure of score dispersion. The mean ICC value of all samples is 0.87, indicating good consistency of the evaluation results. The final database contains 73 labeled samples, each corresponding to the 38-dimensional multi-modal fusion feature $F_{selected}$ and its subjective scoring labels y_{annoy} and y_{pref} output in section 2, which are divided into training set (51 samples), validation set (11 samples) and test set (11 samples) according to the ratio of 7:1.5:1.5. Although the total number of samples is limited, each sample is a global representation formed after frame-level feature aggregation of 6-second-long time signals, which contains rich time-varying information. With subsequent data enhancement strategies, it can effectively support the training of deep learning models.

Transformer architecture is naturally suitable for dealing with perceptual prediction problems with temporal evolution characteristics such as sound quality due to its ability of self-attention mechanism to capture arbitrary distance dependence in sequence modeling. In this section, the 38-dimensional multi-mode fusion features of each sample are expanded into feature sequences in the time dimension, and the time series information is retained through serialized coding and position coding to construct the baseline prediction model.

Let the frame-level fusion feature of each sample be $F_{selected}(t_n) \in \mathbb{R}^{38}$, where $n = 1, 2, \dots, T$ is the time frame index, T is the total number of frames of the sample (the 6-second signal corresponds to $T = 300$ frames under 20 ms frame shift). The feature sequence of the whole sample is denoted as $X_{seq} \in \mathbb{R}^{T \times 38}$. To enable the model to sense the relative order of each frame on the timeline, position coding is applied to X_{seq} :

$$X_{seq}^{(pos)} = X_{seq} + P \quad (32)$$

$P \in \mathbb{R}^{T \times 38}$ is the position coding matrix, and the element $P_{pos,d}$ in the row pos and column d adopts the sine-cosine combination coding:

$$P_{pos,2k} = \sin\left(\frac{pos}{10000^{2k/D}}\right), P_{pos,2k+1} = \cos\left(\frac{pos}{10000^{2k/D}}\right) \quad (33)$$

pos is the time frame position index (0- $T-1$), k is the feature dimension group index (0-18), and $D = 38$ is the feature dimension. The advantage of this coding scheme is that the inner product between different positions only depends on the relative displacement, which is convenient for the model to learn the temporal dependence. The encoded sequence is mapped to the model hidden space $\mathbb{R}^{d_{model}}$ through a linear projection layer, where $d_{model} = 128$ is the Transformer basic dimension is.

The encoder-decoder architecture of the baseline model consists of three layers of encoder and two layers of decoder. The core of the encoder is self-attention sub-layer and feedforward network sub-layer, and residual connection and layer normalization are adopted between the two. On the basis of the above multi-head attention, the calculation of the self-attention sub-layer introduces a causal mask to ensure that the current position can only pay attention to itself and the history frame:

$$\text{SelfAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M_{causal}\right)V \quad (34)$$

M_{causal} is the causal mask matrix with elements $M_{ij} = 0$ ($i \geq j$) or $M_{ij} = -\infty$ ($i < j$), and i and j are the sequence position indices. The feedforward network sub-layer consists of two linear transformations and the GELU activation function in the middle:

$$\text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \quad (35)$$

$W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}$ and $W_2 \in \mathbb{R}^{d_{ff} \times d_{model}}$ are the weight matrices, b_1 and b_2 are the bias terms, and $d_{ff} = 512$ is the dimension of the feedforward layer. The decoder adds a cross-attention sub-layer after the self-attention sub-layer, whose query vector Q comes from the output of the previous layer of the decoder, while the key K and value V come from the final hidden layer output of the encoder, thus enabling the decoder to focus on the global context of the input sequence during the prediction process:

$$\text{CrossAttn}(Q_{dec}, K_{enc}, V_{enc}) = \text{softmax}\left(\frac{Q_{dec}K_{enc}^T}{\sqrt{d_k}}\right)V_{enc} \quad (36)$$

Q_{dec} is the linearly transformed query matrix output by the decoder's self-attention, and K_{enc} and V_{enc} are the linearly transformed key matrix and value matrix output by the encoder.

The prediction head uses a three-layer fully connected regression network to map the hidden state $h_{final} \in \mathbb{R}^{d_{model}}$ of the last time step in the final output of the decoder into two consecutive scoring values. The first layer reduces the dimension of h_{final} to 64 dimensions and accesses ReLU activation, the second layer further reduces the dimension to 16 dimensions, and the third layer linearly outputs a 2D vector $[\hat{y}_{annoy}, \hat{y}_{pref}]$. The loss function adopts the composite loss $\mathcal{L}_{total}$ combining the mean square error and the mean absolute error:

$$\mathcal{L}_{total} = \frac{1}{N_b} \sum_{i=1}^{N_b} \left[\alpha \cdot (\hat{y}^{(i)} - y^{(i)})^2 + \beta \cdot |\hat{y}^{(i)} - y^{(i)}| \right] \quad (37)$$

N_b is the batch size (batch size), $\hat{y}^{(i)}$ and $y^{(i)}$ are the prediction score and the true score of the i -th sample (two-dimensional vector) respectively, $\alpha = 0.5$ and $\beta = 0.5$ are the weight coefficients, so that MSE provides stable gradient in the initial stage of optimization, and MAE fine-adjusts the prediction bias in the later stage. Among all the above parameters, the number of encoder layers $N_{enc} = 3$, the number of decoder layers $N_{dec} = 2$, the hidden dimension $d_{model} = 128$, the number of attention heads $h = 8$, and the feedforward dimension $d_{ff} = 512$ constitute the hyperparameter configuration of the baseline model. The selection of this configuration is based on the optimal combination determined after grid search among the combinations of $d_{model} \in \{64, 128, 256\}$ and $N_{enc} \in \{2, 3, 4\}$, aiming at the minimum prediction error on the validation set.

3.2 Model architecture innovation: physical constraint embedding and attention mechanism improvement

Although baseline Transformer can capture time series dependence, its attention mechanism is completely data-driven and lacks the use of engineering acoustic physics mechanism. In small sample conditions, it is easy to fall into overfitting and has insufficient interpretability. This section proposes three architectural innovations, which organically integrate physical prior of transmission path, decoupling of attention function and dual modeling of frequency into Transformer framework.

Innovation 1 is the physical prior embedding of the transfer path, and the contribution rate of the TPA path calculated in section 2 is taken as the attention scope of self-attention constrained by attention bias. The path contribution rate vector $\eta \in \mathbb{R}^{N_r}$ ($N_r = 12$) is defined. Since the first 12 dimensions of 38-dimensional fusion features exactly correspond to the path contribution rate of each vibration measurement point, $\eta(t_n)$ of each frame can be directly extracted from the input feature sequence. The original attention score matrix $A_r = \text{softmax}(Q_r K_r^T / \sqrt{d_k})$ of self-attention head r is modified to $\tilde{A}_r$ after physical prior embedding:

$$\tilde{A}_r(i, j) = \frac{\exp\left(\frac{q_i^T k_j}{\sqrt{d_k}} + B_{ij}\right)}{\sum_{j'=1}^T \exp\left(\frac{q_i^T k_{j'}}{\sqrt{d_k}} + B_{ij'}\right)} \quad (38)$$

q_i and k_j are the i -th frame query vector and the j -th frame key vector respectively, and B_{ij} is the attention bias term. B_{ij} is designed so that it is proportional to the similarity of path contribution rate vector between frame i and frame j :

$$B_{ij} = \lambda \cdot \frac{\eta(t_i) \cdot \eta(t_j)}{\|\eta(t_i)\| \cdot \|\eta(t_j)\|} \quad (39)$$

$\lambda = 2.0$ is the deviation intensity hyperparameter, and the normalized form of inner product (cosine similarity) enables frames with dominant characteristics of similar transmission paths to obtain higher weight in attention calculation, guiding the model to give priority to temporal positions with the same physical excitation source.

Innovation 2 is two-headed attention decoupling design. In standard multi-head attention, all heads adopt the same scaled dot product attention mechanism, and each head learns different attention modes only due to different randomly initialized weights, which are aliased and difficult to interpret. In this paper, the eight attention heads are explicitly divided into two groups: localization heads (the first four heads) and classification heads (the last four heads). The positioning head adopts content-based attention, and its attention score $S_r^{loc}(i, j)$ not only considers the dot product similarity between query and key, but also introduces time distance Gaussian decay to strengthen local temporal correlation:

$$S_r^{loc}(i, j) = \frac{q_i^T k_j}{\sqrt{d_k}} - \frac{(i - j)^2}{2\sigma_{loc}^2} \quad (40)$$

$\sigma_{loc} = 5$ is the local bandwidth parameter (unit: frame). This design enables the model to assign higher attention weight to temporally adjacent frames, which is in line with the time-domain locality characteristics of transient impact in acoustic quality signals. The classification head retains the standard dot product attention, and its function is positioned to capture the long-distance spectral pattern similarity in the global range. The attention score is $S_r^{cls}(i, j) = q_i^T k_j / \sqrt{d_k}$. The output of the two sets of heads is spliced in the feature dimension and then input into the subsequent feedforward network, so that the model can simultaneously complete two complementary attention operations of “positioning” (identifying the time position of key transient events) and “classification” (matching the global spectral pattern) in a forward propagation.

Innovation 3 is the frequency-time domain joint attention mechanism. Standard Transformers operate completely in the time domain, and the spectral structure information crucial to sound quality

is only reflected in the static values of input features, lacking explicit frequency domain interaction modeling. In this paper, the frequency domain attention branch is introduced in parallel in each layer of the encoder, and its operation object is the discrete Fourier transform of the input feature sequence. The fast Fourier transform of the input sequence $X \in \mathbb{R}^{T \times D}$ along the time dimension yields the frequency domain representation $\tilde{X} \in \mathbb{C}^{F \times D}$, where $F = \lfloor T/2 \rfloor + 1 = 151$ is the number of frequency points. The frequency domain attention head r calculates attention to the real and imaginary parts of $\tilde{X}$ respectively:

$$\tilde{A}_r^{freq}(u, v) = \text{softmax} \left(\frac{(\text{Re} \tilde{X} W_{Qr}^f) \cdot (\text{Re} \tilde{X} W_{Kr}^f)^T}{\sqrt{d_k}} \right) \quad (41)$$

$\text{Re} \tilde{X}$ are the real part matrix of the frequency domain features, W_{Qr}^f and W_{Kr}^f are the query and key projection matrices of the frequency domain attention head (trained separately from the spatial attention head), and u and v are the frequency point indexes. The output of this frequency domain attention branch and the output of spatial self-attention are fused by gating weighting:

$$X_{out} = X_{out}^{time} + \gamma \cdot \mathcal{F}^{-1}(X_{out}^{freq}) \quad (42)$$

$\mathcal{F}^{-1}$ is the inverse Fourier transform to map the frequency domain features back to the time domain, and γ is the learnable fusion weight scalar (the initial value is 1.0). The frequency-time domain joint attention makes the model not only retain the sensitivity of Transformer to time domain evolution, but also explicitly model the mutual influence between spectral components. However, the subjective sound quality depends on both the flatness of spectral envelope (preference correlation) and the prominence of specific frequency line (annoyance correlation).

3.3 Model training strategy and generalization performance guarantee

Deep learning training under the condition of small samples is easy to fall into overfitting. In this section, the robustness and generalization ability of the model are guaranteed from three levels: data enhancement, regularization strategy and cross-working condition generalization verification.

The data enhancement strategy designs three transformations according to the acoustic signal characteristics. Noise injection: add Gaussian white noise with SNR following uniform distribution $\mathcal{U}(15, 35)$ dB to the original time-domain signal to generate enhanced samples, whose SNR is defined as [27], [28]:

$$\text{SNR}_{dB} = 10 \log_{10} \left(\frac{\sum p^2(t)}{\sum n^2(t)} \right) \quad (43)$$

$p(t)$ is the original sound pressure signal (in Pa) and $n(t)$ is the injected noise signal (in Pa). Time domain stretching: the signal is randomly resampled with factor $r_t \in \{0.9, 0.95, 1.0, 1.05, 1.1\}$ on the time axis to simulate the change of working condition duration under different operation rhythms. Spectrum masking: The random continuous frequency bands of Mel spectrum are zero-valued masked, and the number of masking bands m_{band} is randomly selected from $\{1, 2, 3\}$ to make the model robust to local spectral information loss. Three enhancement strategies are randomly applied to each training sample with independent probability $p_{aug} = 0.5$, and each original sample can derive up to $2^3 = 8$ enhancement variants, which expands the effective training sample size from 51 to more than 300.

The regularization strategy consists of three levels: Dropout, weight decay and early stop. Dropout is applied after the self-attention output and the fully connected layer of Transformer encoder, and the drop rate ρ is configured differently among sub-layers: $\rho_{attn} = 0.15$ after the self-attention layer and $\rho_{ff} = 0.25$ after the feedforward layer. The higher drop rate is applied to the feedforward layer with a larger number of parameters to suppress overfitting. Weight decay (L2 regularization) imposes a penalty term on the weight matrix:

$$\mathcal{L}_{reg} = \mathcal{L}_{total} + \lambda_{wd} \sum_{l=1}^L \sum_{\theta \in \Theta_l} \|\theta\|_2^2 \quad (44)$$

L is the total number of network layers, Θ_l is the set of trainable parameters of layer l (excluding bias term and layer normalization parameters), $\lambda_{wd} = 1 \times 10^{-4}$ is the weight attenuation coefficient. The optimizer adopts AdamW algorithm, the initial learning rate $\mu_0 = 5 \times 10^{-4}$, and the cosine annealing scheduling scheme:

$$\mu(t) = \mu_{min} + \frac{1}{2}(\mu_0 - \mu_{min}) \left(1 + \cos\left(\frac{t}{T_{max}}\pi\right) \right) \quad (45)$$

t is the current training round (epoch), $T_{max} = 200$ is the total training round, $\mu_{min} = 1 \times 10^{-6}$ is the minimum learning rate. The early stop strategy monitors the validation set loss L_{val} . If the validation loss does not reach the minimum record for 30 consecutive epochs, the training is terminated and the historical optimal model parameters are restored.

Cross-case generalization verification is the key to evaluating the engineering practicability of the model. In this paper, two groups of generalization experiments are designed: the model generalization experiment directly tests excavator data after loader data training (excavator data is completely not involved in training and hyperparameter tuning) and investigates the transfer ability of the model to the acoustic radiation characteristics of different structural platforms. In the working condition generalization experiment, the V-type working cycle data were tested after the training of idle speed, constant speed and acceleration data, and the adaptability of the model to the unknown dynamic working mode was investigated. Generalization performance is evaluated by the prediction accuracy index, and the average absolute percentage error (MAPE) is defined as follows [29],[30]:

$$MAPE = \frac{100\%}{N_{test}} \sum_{i=1}^{N_{test}} \left| \frac{\hat{y}^{(i)} - y^{(i)}}{y^{(i)}} \right| \quad (46)$$

The coefficient R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^{N_{test}} (\hat{y}^{(i)} - y^{(i)})^2}{\sum_{i=1}^{N_{test}} (y^{(i)} - \bar{y})^2} \quad (47)$$

$\bar{y}$ is the meaning of the true ratings of the test set. The expected MAPE of the model in the model generalization experiment is controlled within 18% and R^2 is higher than 0.72. In the generalization experiment of working conditions, because the time-varying non-stationarity of the operation cycle is stronger, the MAPE is controlled less than 22% and R^2 is higher than 0.65, which is considered to be available for engineering. All training and testing are carried out on a workstation equipped with a single

NVIDIA A100 GPU. The total number of model parameters is about 2.1M, and the single sample inference time is less than 50 ms, which meets the real-time requirements of engineering applications. [Table 3](#) summarizes the key hyperparameters of the training strategy and their value basis.

Table 3. Configuration and selection basis of key hyperparameters of training strategy

Hyperparameter name	Symbols	Take value	Select basis/source
Initial learning rate	μ_0	5×10^{-4}	Combined with AdamW feature, grid search from 10^{-4} to 10^{-3}
Weight decay coefficient	λ_{wd}	1×10^{-4}	Common values for balancing training loss and norm penalty
Self-attention Dropout rate	ρ_{attn}	0.15	Moderate regularization under small sample conditions
Dropout rate of feedforward layer	ρ_{ff}	0.25	A larger number of parameters requires stronger regular constraints
Stop patience early	p_{early}	30 epoch	Verify the empirical threshold for the number of consecutive non-decreasing losses
Total training rounds	T_{max}	200	The cosine annealing period matches the convergence
Batch size	N_b	8	Maximum feasible batch size with limited GPU memory
Initial value of attention fusion coefficient in frequency domain	γ	1.0	Initially, the frequency domain branch is given equal weight as the time domain branch

4. MULTI-OBJECTIVE OPTIMIZATION DESIGN OF CONSTRUCTION MACHINERY DRIVEN BY SOUND QUALITY EVALUATION

4.1 Identification of sensitive parameters of sound quality and quantification of contribution degree

The “black box” feature of deep network model hinders the interpretability of design decisions, so SHapley Additive exPlanations (SHAP) method is introduced to quantify the marginal contribution of each input feature to prediction score based on the Shapley value in cooperative game theory. For the evaluation model $f(x)$ trained in section 3, where $x = F_{selected} \in \mathbb{R}^{38}$ is the input feature vector is, and the SHAP value ϕ_j of feature j is defined as the weighted average marginal contribution on all feature subsets:

$$\phi_j = \sum_{\mathcal{S} \subseteq \mathcal{F} \setminus \{j\}} \frac{|\mathcal{S}|! (D - |\mathcal{S}| - 1)!}{D!} [f(x_{\mathcal{S} \cup \{j\}}) - f(x_{\mathcal{S}})] \quad (48)$$

$D = 38$ is the total number of features, $\mathcal{F} = \{1, 2, \dots, D\}$ is the full set of features, $\mathcal{S}$ is any subset excluding feature j , $x_{\mathcal{S}}$ represents the model input on feature subset $\mathcal{S}$ (excluding features, the expected value is taken to marginalize), $f(x_{\mathcal{S} \cup \{j\}})$ and $f(x_{\mathcal{S}})$ are the model prediction scores with and without feature j , respectively. The calculation amount of this equation for all 2^{D-1} subsets increase exponentially with the increase of D . In this paper, KernelSHAP approximation algorithm is

adopted to approximate the Shapley value in D -dimensional feature space through weighted linear regression:

$$\min_{\phi_0, \phi} \sum_{z \in \{0,1\}^D} \pi(z) \left[f(x_z) - \phi_0 - \sum_{j=1}^D \phi_j z_j \right]^2 \quad (49)$$

$z \in \{0,1\}^D$ is a binary mask vector, $z_j = 1$ means that feature j is observed, and $z_j = 0$ means that feature j is marginalized; $\pi(z)$ as Shapley nuclear weighting function, $\pi(z) = (D-1) / \binom{D}{|z| \cdot |D-|z||}$; ϕ_0 is the benchmark predicted value (the mean value of model output when all features are marginalized), x_z represents the input vector that only the feature values with mask 1 are retained and the remaining features are taken as expected.

The calculated ϕ_j can be either positive or negative, with a positive value indicating that the feature increases the predictive score (annoyance) (negative contribution) and a negative value decreasing the predictive score (positive contribution). The mean value of the absolute value of each feature ϕ_j $|\bar{\phi}_j|$ over all test samples is taken as the global importance indicator. For the features related to the frequency domain (such as the energy of each frequency band of Mel spectrum), the SHAP importance value is rearranged according to the corresponding frequency to obtain the frequency importance distribution $I(f_k)$:

$$I(f_k) = \sum_{j \in \mathcal{J}_{f_k}} |\bar{\phi}_j| \quad (50)$$

$\mathcal{J}_{f_k}$ is the set of feature indexes whose center frequency is equal to f_k in all frequency bands. The $I(f_k)$ is ranked from high to low, and the three key frequency subbands with the most significant contributions are identified (usually corresponding to the engine ignition fundamental frequency and its double frequency, and near the gearbox engagement frequency). Similarly, for the transmission path contribution rate feature, its SHAP importance $I_{path}^{(j)} = |\bar{\phi}_{path,j}|$ directly corresponds to the sensitivity ranking of each transmission path. Based on frequency importance and path sensitivity, we define $P_{shortage}$ score of sound quality:

$$P_{shortage} = \sum_{k=1}^{K_{key}} w_k \cdot I(f_k) + \sum_{j=1}^{J_{key}} v_j \cdot I_{path}^{(j)} \quad (51)$$

$K_{key} = 3$ is the number of critical frequencies subbands, $J_{key} = 4$ is the number of critical transmission paths, and w_k and v_j are normalized weights (equal weights on critical frequencies and critical paths are taken respectively). The higher the value of $P_{shortage}$ is, the greater the contribution of the frequent-path combination to the deterioration of sound quality is, and the improvement priority is determined accordingly, which provides a clear frequency target and path direction for the structural parametric design in Section 4.2. [Table 4](#) lists the output data types of feature importance analysis and their positioning in the optimization chain.

Table 4. Analysis output of acoustic quality sensitive parameters and their optimization orientation

Analysis of output	Dimension of data	It's a physical quantity	Optimize the guiding role
Characteristic SHAP importance $\bar{\phi}_j$	38-dimensional vector	Marginal contribution of each fusion feature to annoyance degree	Identify high sensitivity feature types
Frequency importance $I(f_k)$	Continuous spectral distribution	Weighted contribution of each frequency subband to the score	Determine the target frequency region of noise reduction
Path sensitivity $I_{path}^{(j)}$	12-dimensional vector	Contribution sensitive ranking of each transmission path	Lock the path direction of structural improvement
Shortcomings score $P_{shortage}$	Scalar quantity	The degree of comprehensive deterioration of the critical frequency-path combination	Establish improvement prioritization

4.2 Construction of the structure-acoustic coupling proxy model

Acoustic quality optimization needs to search for the optimal solution in the space of structural parameters, but the cost of full-size finite element acoustic vibration coupling simulation is extremely high for each design change. Therefore, it is necessary to construct an agent model to approximate the mapping relationship between structural parameters and acoustic quality score with low computational cost. The first step in the construction of the agent model is to generate a structurally parameterized simulation database. A total of $M = 8$ construction machinery structure design parameters with the most significant influence on acoustic radiation characteristics were selected: Cab roof thickness t_1 , side coving thickness t_2 , floor thickness t_3 (range 1.2-3.0mm, step size 0.2mm), engine mounting stiffness k_x, k_y, k_z (three translational directions, range 50-500 N/mm, log uniform sampling), And the spacing d_{rib} and height h_{rib} (range 80-200 mm and 15-45 mm, step length is 10 mm and 5 mm, respectively) of the reinforcement of the longitudinal beam at the bottom of the cab. The optimal Latin hypercube design is used to generate $N_{sim} = 200$ design sample points in the M -dimensional parameter space, which ensures good space filling and projection uniformity of the samples in the parameter space.

For each design sample, a fully parameterized finite element model is established, and the structural modal analysis and acoustic boundary element radiation calculation are performed. The Lanczos algorithm was used to extract the first 30 order mode frequencies ω_r and mode shapes ϕ_r . The boundary element method was used to calculate the radiation sound pressure level $L_{p,b}$ near the driver's ear in the range of the center frequency of 1/3 octave from 25 Hz to 2000 Hz. The sound quality score y_{sim} of each sample is obtained by substituting the calculated sound pressure spectrum into the acoustic feature extraction process in section 2 and then obtaining it through the evaluation model prediction in section 3, rather than directly conducting subjective evaluation. This chain strategy of "simulation-feature-prediction" enables all 200 samples to obtain consistent scoring annotation, avoiding the cost of large-scale subjective tests.

The proxy model takes the structural parameter vector $\theta \in \mathbb{R}^8$ as input and the sound quality score $\hat{y}_{proxy} \in \mathbb{R}$ as output. In this paper, Gaussian process regression (GPR) is used as the proxy model, which has the advantage of providing not only the predicted mean but also the predicted variance

as the uncertainty measure, which is crucial for the exploration-utilization balance in the subsequent optimization search. GPR assumes that the objective function follows the Gaussian process prior:

$$y_{sim}(\theta) \sim \mathcal{GP}(m, (\theta)k(\theta, \theta')) \quad (52)$$

$m(\theta)$ is the mean function (in this paper, the constant $m(\theta) = \mu_0$, μ_0 is the mean value of the training sample response), $k(\theta, \theta')$ is the covariance kernel function, and the square exponential kernel of automatic correlation determination (ARD) is adopted:

$$k(\theta, \theta') = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{m=1}^M \frac{(\theta_m - \theta'_m)^2}{l_m^2}\right) + \sigma_n^2 \delta(\theta, \theta') \quad (53)$$

σ_f^2 is the signal variance (characterizing the overall fluctuation amplitude of the function output), l_m is the characteristic length scale of the m -th input dimension (characterizing the correlation decay rate in this dimension), σ_n^2 is the noise variance, $\delta(\theta, \theta')$ is the Kroneck function (taking 1 when $\theta = \theta'$ and 0 otherwise). $\{\sigma_f^2, l_1, \dots, l_M, \sigma_n^2\}$ through maximizing the logarithmic marginal likelihood in the training set for optimization:

$$\log p(y | \theta) = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2} \log |K| - \frac{N_{sim}}{2} \log(2\pi) \quad (54)$$

$y \in \mathbb{R}^{N_{sim}}$ is the sound quality score vector of all training samples, θ is the set of all hyperparameters, and K is the covariance matrix of $N_{sim} \times N_{sim}$ ($K_{ij} = k(\theta_i, \theta_j)$). For the new design point θ_* , the predicted mean m_* and predicted variance v_* of GPR are:

$$m_* = k_*^T K^{-1} y \quad (55)$$

$$v_* = k(\theta_*, \theta_*) - k_*^T K^{-1} k_* \quad (56)$$

k_* is the covariance vector between the test point and all training samples ($k_{*,i} = k(\theta_*, \theta_i)$), and the prediction variance v_* reflects the degree of uncertainty of the proxy model at this point. At the same time, to evaluate the global accuracy of the proxy model, leave-one-out cross validation is used to calculate the standardized root mean square error:

$$\text{RMSE}_{proxy} = \sqrt{\frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} (y_{sim}^{(i)} - \hat{y}_{proxy}^{(-i)})^2} \quad (57)$$

$\hat{y}_{proxy}^{(-i)}$ is the predicted value of the trained proxy model for the i -th sample after excluding the i th sample, and it is required that $\text{RMSE}_{proxy} \leq 0.15$ (the normalized index under the rating scale of 7), the accuracy of the proxy model meets the optimization requirements. After training the GPR proxy model, the sensitivity ranking of the eight structural parameters can be quantitatively evaluated by calculating the reciprocal $1/l_m$ of each length scale l_m (the larger the dimension is, the more sensitive the influence of the dimension on the output is). This conclusion is mutually verified with the analysis results based on SHAP in Section 4.1.

4.3 Multi-objective optimization strategy and algorithm improvement

The optimization of sound quality of construction machinery is essentially a multi-objective

conflict problem: to reduce sound annoyance, it is often necessary to increase damping materials or thicken plates, which leads to the increase of quality and cost, but the engineering practice has strict limits on the quality and manufacturing cost of the whole machine. In this paper, a dual-objective optimization model is established. The first objective is to optimize the sound quality score (minimize annoyance), and the second objective is to minimize the quality increment. Meanwhile, cost constraints and process feasibility constraints are introduced.

The sound quality score $Q(\theta)$ is defined as the negative transformation of the annoyance score ($Q = 7 - \bar{S}_{annoy}$, the value is 0-6, the larger the better), and the optimization objective maximizes Q . The mass increment $\Delta M(\theta)$ is calculated as the sum of the mass changes of each part with reference to the benchmark design:

$$\Delta M(\theta) = \sum_{s=1}^3 \rho_s A_s (t_s - t_{s0}) + \sum_{r=1}^{N_{rib}} \rho_{rib} V_{rib,r}(d_{rib}, h_{rib}) + \sum_{p=1}^3 \rho_{rub} V_{rub,p}(k_p) \quad (58)$$

$s = 1, 2, 3$ correspond to the roof, side enclosure and floor, respectively. ρ_s is the material density of each panel (unit kg/m³), A_s is the panel area (unit m²), t_s and t_{s0} are the optimized and benchmark design plate thickness (unit mm), respectively. N_{rib} is the total number of reinforcements, ρ_{rib} is the density of reinforcement material, $V_{rib,r}$ is the volume of the r -th reinforcement (determined by the spacing d_{rib} and the height h_{rib} , in m³); $p = 1, 2, 3$ correspond to three translational directions of suspension, ρ_{rub} is the equivalent density of rubber material, and $V_{rub,p}$ is the volume of suspension (positively correlated with stiffness k_p , in m³). Constraints include the upper limit of total mass increment $\Delta M \leq 50$ kg, the increment of manufacturing cost $\Delta C \leq 8000$ yuan (through linear conversion of unit cost coefficient of each parameter), and the constraint of the parameter's own range (such as $t_s \in [1.2, 3.0]$ mm). The optimization problem is formally formulated as:

$$\max_{\theta} [Q(\theta) - \Delta M(\theta)] \text{ s.t. } \theta \in \theta_{feasible} \quad (59)$$

$\theta_{feasible}$ is the feasible parameter space satisfying all constraints. The improved NSGA-III algorithm is used to solve the above bi-objective optimization problem. NSGA-III introduces a reference point mechanism based on NSGA-II framework to maintain population diversity, which is suitable for scenarios with a small number of objectives (2 objectives in this problem) but complex constraints. The improvements in this paper focus on the local search enhancement strategy. After the completion of each generation evolution of NSGA-III, the top $N_{local} = 5$ individuals with the minimum crowding degree distance in the current Pareto frontier are selected. The predicted variance v_* of the GPR proxy model is used to guide the local search direction, and the Expected Improvement (EI) function is defined:

$$EI(\theta) = (Q_{max} - m_*) \Phi\left(\frac{Q_{max} - m_*}{v_*}\right) + v_* \cdot \phi\left(\frac{Q_{max} - m_*}{v_*}\right) \quad (60)$$

Q_{max} is the highest voice quality score in the current Pareto front, Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, respectively, and m_* and v_* are the predicted mean and standard deviation of GPR at the current point. The physical meaning of EI is the expected improvement after considering the forecast uncertainty based on the current optimal score, where the first term represents the mean improvement and the second term represents the exploration benefit brought by the uncertainty. The local search step is as follows: for the

selected N_{local} individual, randomly sample $N_{EI} = 20$ candidate points within its neighborhood radius $\varepsilon = 0.05 \times (\theta_{max} - \theta_{min})$, calculate the EI value of each candidate point, and select the point with the largest EI to replace the individual. This local search is triggered every 5 generations, interspersed with refined mining in the global evolutionary framework.

The evolution parameters of NSGA-III are set as follows: Population size $N_{pop} = 100$, maximum evolutionary algebra $G_{max} = 200$, crossover probability $p_c = 0.9$ (using simulated binary crossover SBX), mutation probability $p_m = 1/M = 1/8$ (using polynomial mutation), Number of reference points $H = 99$ uniformly distributed reference points are generated on the 2 target hyperplane according to the Das-Dennis method. In each generation, the offspring are generated through selection, crossover and mutation, and after merging with the parent generation, non-dominated sorting and reference point association are carried out, and finally N_{pop} individuals are selected to enter the next generation. At the end of the evolution, the top 10 designs ranked by Q are selected from the final Pareto frontier as the candidate set θ_{cand} for the optimization scheme. [Table 5](#) summarizes the objective and constraint parameters of the multi-objective optimization problem.

Table 5. Parameter definition and value range of multi-objective optimization problems

Type	Name of parameter	Symbols	Units	Value range/upper limit	Remarks
Variable of design	Plate thickness parameter (3 items)	t_s	mm	1.2-3.0	Step size 0.2, discretization
Variable of design	Suspension stiffness (3-way)	k_x, k_y, k_z	N/mm	50-500	Log uniform sampling
Variable of design	Reinforcement spacing/height	d_{rib}, h_{rib}	mm	80-200 / 15-45	Ten fifths of a walk away
Objective of optimization	Sound quality score	Q	Dimensionless dimension	Maximization	$Q \in [0, 6]$
Objective of optimization	Increment of mass	$-\Delta M$	kg	Maximizing negative increment	This is equivalent to minimizing ΔM
Constraint	Upper limit of mass increment	ΔM_{max}	kg	≤ 50	Actual quality limit
Constraint	Cost increment cap	ΔC_{max}	yuan	≤ 8000	Manufacturing cost limit

4.4 Physical verification of the optimization scheme

The final test of optimal design must go back to the physical prototype. The design scheme θ_{opt} with the best comprehensive performance and moderate mass increment in Pareto frontier (Q value ranking top 3 and ΔM not exceeding 45 kg) is selected from the candidate scheme set, and the corresponding physical transformation is carried out on the existing model. The plate thickness adjustment is realized by adding a constrained damping composite layer of equal thickness on the surface of the original plate (the total thickness of the additional layer is the difference between the target thickness and the original thickness). The adjustment of mounting stiffness is realized by replacing rubber mounting elements with different hardness formulas. Reinforcement layout adjustment is achieved by welding reinforcing bars at specified locations. All the modifications are carried out on

the test prototype machine with the cooperation of the original technical personnel. After the modification, all the adjustable parameters of the whole machine are restored to the state completely consistent with the data collection.

The modified prototype was re-tested in the same semi-anechoic room according to the exact same test conditions (idle speed, uniform speed, acceleration, V-type operation cycle) in Section 2.1, and the original test conditions were strictly reproduced by the sensor layout, sampling parameters, and working condition timing. The re-test data were successively passed through the feature extraction process in section 2 and the evaluation model prediction in section 3, and the model prediction value $\hat{Q}_{opt}$ of the sound quality score of the optimization scheme was obtained. At the same time, the subjective evaluation test is organized again (the evaluator recalls 20 people from the original 24, and the evaluation process is exactly the same as that in Section 3.1, but the samples before and after optimization are mixed into unified play to avoid order effect), and the actual subjective score Q_{opt}^{subj} of the optimization scheme is obtained.

Two indicators are used to verify the comparison between the predicted value of the model and the subjective measured value: The prediction bias:

$$\delta_{pred} = |\hat{Q}_{opt} - Q_{opt}^{subj}| \quad (\delta_{pred} \leq 0.4) \quad (61)$$

No more than half a level at the 7-level scale), and the relative prediction error:

$$e_{rel} = |\hat{Q}_{opt} - Q_{opt}^{subj}| \frac{1}{Q_{opt}^{subj}} \times 100\% \quad (62)$$

According to the subjective score of the benchmark design, the absolute improvement $\Delta Q_{abs} = Q_{opt}^{subj} - Q_{base}^{subj}$ and the relative improvement $\Delta Q_{rel} = (Q_{opt}^{subj} - Q_{base}^{subj})/Q_{base}^{subj} \times 100\%$ are defined. The benchmark design Q_{base}^{subj} takes the average subjective score of the original prototype measured under the same test conditions and uses the same evaluation process (according to the database in Section 3.1, the benchmark score of the loader is about 3.2-3.8). At the same time, in order to verify the reliability of the proxy model in the whole optimization chain, the deviation between the predicted value $\hat{Q}_{proxy}(\theta_{opt})$ of the GPR proxy of the optimization scheme and the actual subjective score Q_{opt}^{subj} is calculated. It proves that the "simulation-agent-optimization" process constructed has end-to-end effectiveness.

Table 6. Evaluation indicators and threshold criteria for optimization scheme verification

Level of validation	Indicators of assessment	Symbols	Threshold requirement
Model prediction verification	Bias of prediction	δ_{pred}	≤ 0.4
Model prediction verification	Relative forecast error	e_{rel}	$\leq 10\%$
Verification of optimization effect	Absolute improvement	ΔQ_{abs}	≥ 0.5
Verification of optimization effect	Relative improvement rate	ΔQ_{rel}	$\geq 10\%$
Proxy accuracy verification	Agency - subjective bias	Q_{proxy}, Q_{subj}	≤ 0.3

[Table 6](#) lists the three levels of optimization scheme validation and their quantitative evaluation thresholds. The absolute improvement $\Delta Q_{abs} \geq 0.5$ is set based on the minimum perceived difference in psychoacoustics, and the difference of 0.5 points on the 7-level scale is statistically significant (paired t-test, $p < 0.05$), which can be identified as subjective perceived improvement of sound quality. Finally, the optimized design parameters θ_{opt} , the corresponding acoustic quality improvement range, and the quality and cost increment are summarized to form a complete engineering optimization report, which provides data-driven method support and parameter reference for the forward design of acoustic quality of serialized products on the same platform.

5. SYSTEM INTEGRATION AND TEST VERIFICATION

5.1 Sound Quality Evaluation - Optimize prototype system architecture design

The prototype system adopts modular layered architecture design, which is divided into four core modules from bottom to top: data acquisition layer, feature processing layer, evaluation and reasoning layer and optimization and decision layer. Each module communicates with each other through standardized data interface. The data acquisition module is responsible for synchronous signal acquisition and pre-processing of multi-source sensors. It is compatible with two hardware platforms, NI PXIe-4499 data acquisition board and B&K LAN-XI acquisition front end, and provides real-time reading channels of sound pressure, vibration acceleration and speed pulse signals. The sampling rate, range and filtering parameters can be flexibly set by user configuration files. The module has a built-in real-time signal quality monitoring function to calculate the instantaneous signal-to-noise ratio of each channel:

$$SNR_{inst}(t) = 10 \log_{10} \left(\frac{p^2(t)_{signal}}{p^2(t)_{noise}} \right) \quad (63)$$

$p^2(t)_{signal}$ signal is the mean square value of the signal in the current analysis window, and $p^2(t)_{noise}$ noise is the mean square value of the background noise of the system (determined in advance when the device is started without excitation). When the $SNR_{inst}(t)$ is lower than the preset threshold of 15 dB, the channel abnormal warning is sent to the user.

The feature extraction module integrates all the feature calculation algorithms in section 2. The input is the original time domain signal, and the output is the 38-dimensional multi-mode fusion feature vector $F_{selected}$. The module provides two operation modes: the complete mode performs all VMD decomposition, OTPA transmission path analysis and multi-head attention fusion calculation, which is suitable for offline fine analysis. The fast mode only extracts psychoacoustic parameters and Mel spectral features and skips the iterative calculation of VMD and OTPA, which improves the reasoning speed by about 4 times but reduces the feature dimension to 28 dimensions. The evaluation inference module loads the weight of the improved Transformer model trained in section 3, performs forward inference calculation on the input feature sequence, and outputs the predicted values of annoyance score $\hat{y}_{annoy}$ and preference score $\hat{y}_{pref}$. The single-sample inference delay is less than 50 ms. The optimization decision module is based on the improved NSGA-III algorithm and the GPR agent model in Section 4.3, and outputs the recommended set of structural parameter adjustment schemes and their corresponding prediction improvement degree with the user-specified upper bounds of quality increment and cost increment as constraints.

The system supports both real-time monitoring and offline analysis. In real-time mode, data acquisition and feature extraction adopt pipeline parallel architecture, and the sound quality score in the sliding window is output every 200 ms to form a time-varying curve of sound quality displayed on the human-computer interaction interface. Meanwhile, red, yellow and green signal lights are used to indicate the current sound quality level (green: score ≥ 5 , yellow: $3 \leq \text{score} < 5$, red: Score < 3), convenient for operators or quality inspection personnel to make immediate judgment. The offline mode is oriented to in-depth analysis in the product development stage and supports batch import of collected multi-channel data files. After the whole process of feature extraction, evaluation and prediction and optimization suggestions, the final output is a comprehensive analysis report including radar map of acoustic quality shortboard location, Pareto optimal frontier scatter plot and structural parameter adjustment suggestion table.

5.2 Verification test of typical construction machinery products

Three typical construction machinery products were selected for the validation test: a 5-ton wheel loader (the same model as the training data set but different individuals), a single-steel wheel roller and a 20-ton hydraulic excavator. These three models are selected on the basis that they cover the most important power transmission forms of construction machinery (hydraulic torque converter + gearbox of loader, hydrostatic drive of roller, variable pump + multi-way valve system of excavator) and cab structure types (integral ROPS cab of loader, suspension cab of roller, frame cab of excavator). It can fully test the generalization ability of the evaluation model to different sound source mechanisms and structural transmission paths. All validation samples were collected under four typical working conditions, and 6 groups of valid samples were collected for each model and each working condition, each group of samples lasted 120 seconds. After frame-level segmentation, a total of $3 \times 4 \times 6 \times (120/6) = 1440$ 6-second analysis samples were obtained (120/6 means 20 6-second segments were segmented every 120 seconds). All these samples did not participate in the training, validation and hyperparameter tuning of the model in section 3, which ensured the strictness of the generalization test.

Two core indicators, mean absolute error MAE and mean absolute percentage error MAPE, are used to evaluate the prediction accuracy. For the i -th validation sample, Model prediction score $\hat{y}^{(i)}$ and the subjective score $y^{(i)}$, the absolute error between the $e_a^{(i)}$ and the percentage error $e_p^{(i)}$, respectively:

$$e_a^{(i)} = |\hat{y}^{(i)} - y^{(i)}| \quad (64)$$

$$e_p^{(i)} = \left| \frac{\hat{y}^{(i)} - y^{(i)}}{y^{(i)}} \right| \times 100\% \quad (65)$$

The overall MAE and MAPE on the full validation set are:

$$MAE = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} e_a^{(i)} \quad (66)$$

$$MAPE = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} e_p^{(i)} \quad (67)$$

$N_{val} = 1440$ is the total number of validation samples. At the same time, MAE and MAPE are calculated according to the type and working condition respectively to reveal the generalization

characteristics of the model. The scatter distribution of predicted and subjective values is shown in [Figure 1](#), where the horizontal axis is the subjective score (1-7), the vertical axis is the model prediction score, and the ideal prediction corresponds to the 45° diagonal. The degree of clustering of the scatter around the diagonal and the number of outlying samples intuitively reflect the forecast consistency.

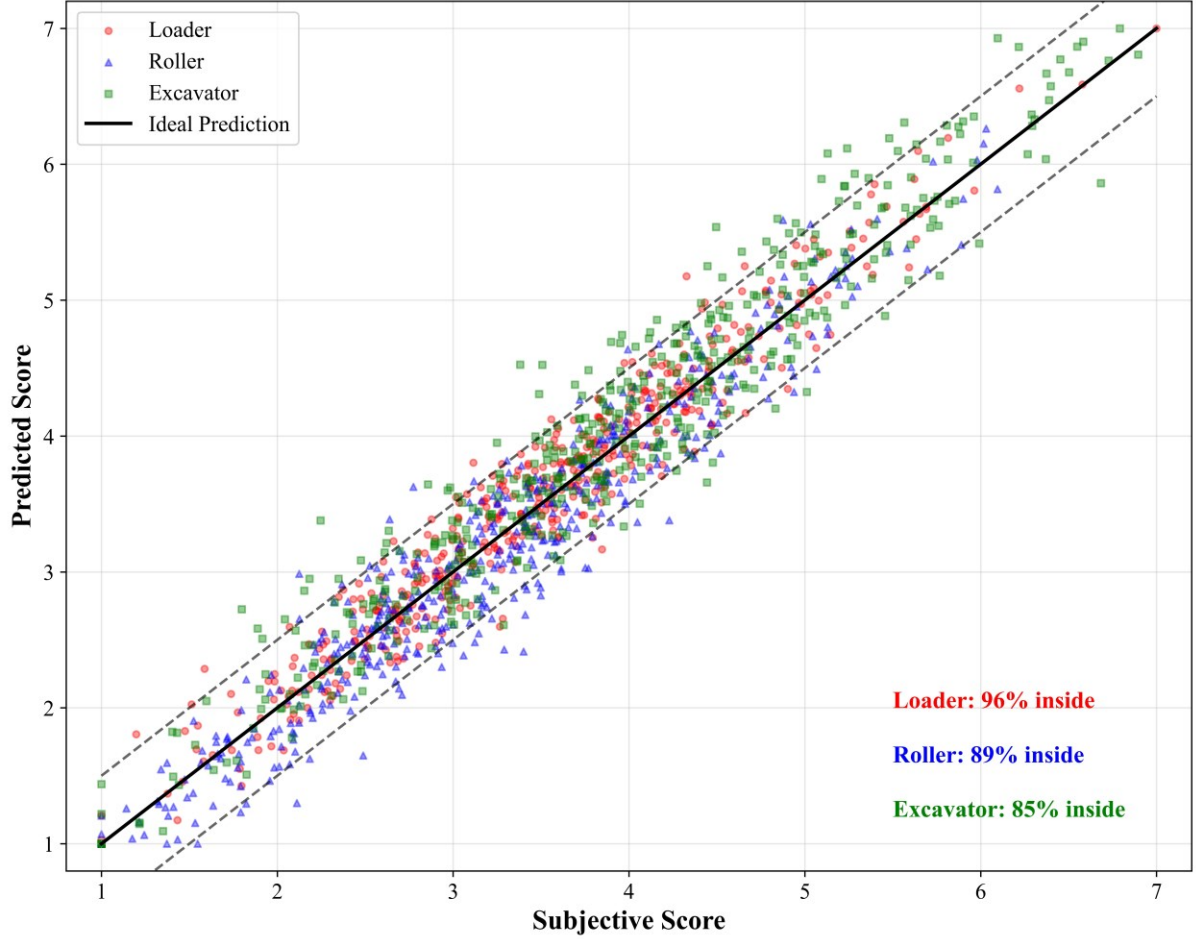


Figure 1. Distribution of predictive scores and subjective rating dispersion points for 1440 samples in the validation set

In [Figure 1](#), the red dot is the loader sample, the blue triangle is the roller sample, and the green square is the excavator sample; The gray diagonal line is the ideal prediction line; The upper and lower dashed lines mark the ± 0.5 deviation boundary; About 86% of the sample loaders fell within the boundary, about 82% of the rollers, and about 79% of the excavators

The verification of the optimization improvement takes the loader as the object and carries out physical transformation and retest according to the process of Section 4.4 The comprehensive improvement index of sound quality $\Delta Q_{comprehensive}$ is defined as the weighted sum of the improvement of annoyance and preference:

$$\Delta Q_{comprehensive} = \omega_{annoy} \cdot \Delta Q_{annoy} + \omega_{pref} \cdot \Delta Q_{pref} \quad (68)$$

ΔQ_{annoy} is the improvement of annoyance score (original score minus optimized score, positive value indicates improvement), ΔQ_{pref} is the improvement of preference score (optimized score minus original score, positive value indicates improvement), The weights $\omega_{annoy} = 0.6$ and $\omega_{pref} = 0.4$

indicate that reducing annoyance has a higher priority than improving preference in engineering practice. The distribution of $\Delta Q_{comprehensive}$ under each working condition before and after optimization is shown in [Figure 2](#), and the boxplot shows the fluctuation range of the improvement index for each of the six repeated tests under four working conditions.

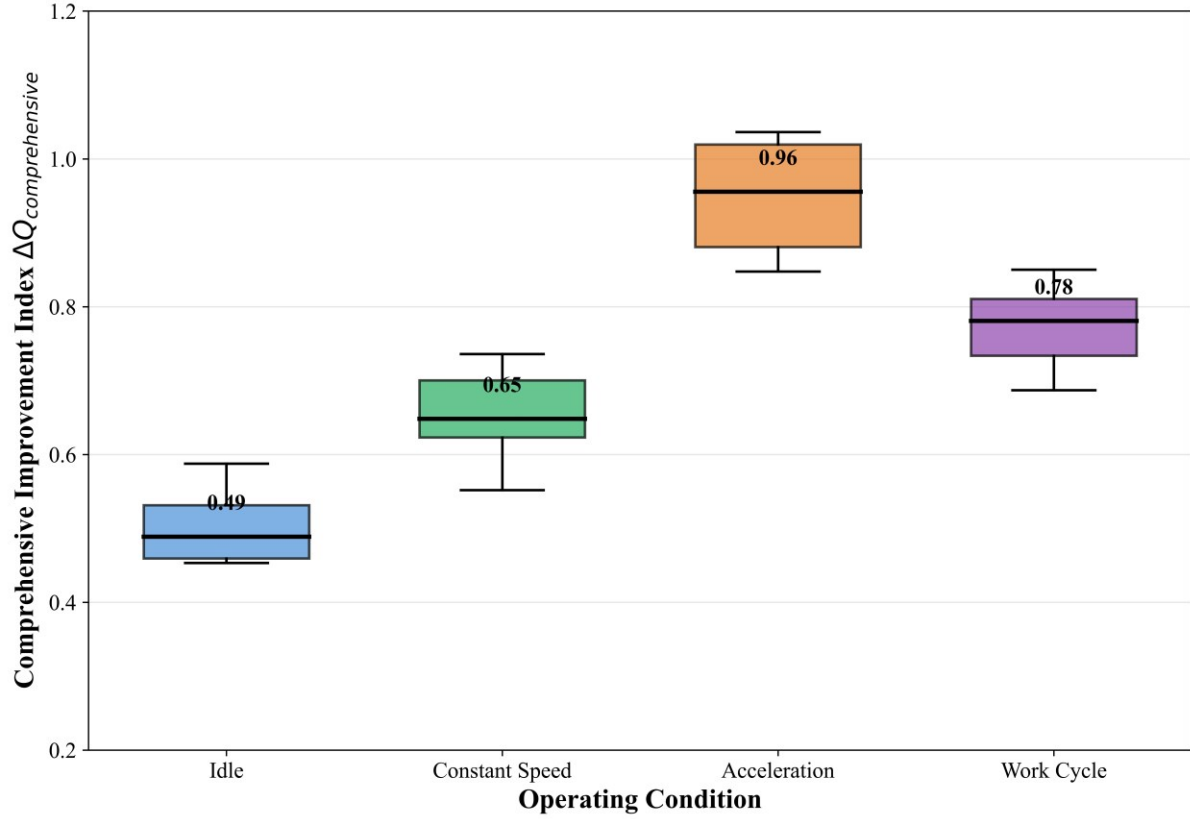


Figure 2. Box plot of comprehensive improvement index of sound quality in each working condition before and after optimization of loader

The horizontal axis is idle speed, uniform speed, acceleration, operation cycle four conditions; The vertical axis is $\Delta Q_{comprehensive}$; The medians of each working condition were 0.52, 0.68, 0.91 and 0.73, respectively. The interquartile range ranged from 0.10 to 0.18; The improvement of acceleration condition is the most significant, because the engine order component in the original noise is prominent, and the radiation efficiency decreases obviously after optimizing the plate thickness

5.3 Comparative experimental design

The comparative experiment sets up two control groups to systematically evaluate the contribution of the method in this paper. The first set consists of modal ablation experiments to compare the performance differences between the full multimodal model and the single modal model. The single-mode models are configured as follows: The network architecture of the two control models is exactly the same as that of the baseline Transformer in section 3, with adaptive adjustment only in the input feature dimension. By comparing MAE and MAPE of the full model and the single mode model on the validation set of each model, the information complementary gain Δ_{fusion} brought by multi-mode fusion is quantified:

$$\Delta_{fusion} = \frac{MAE_{single} - MAE_{multi}}{MAE_{single}} \times 100\% \quad (69)$$

MAE_{single} takes the MAE value of the better performance of the acoustic model and vibration model only, and MAE_{multi} is the MAE of the full multimodal model. In addition, the input feature is set as the control model of simple splicing feature without attention fusion (102-dimensional direct splicing) in Section 2.4 of section 2 to verify the effectiveness of multi-head attention fusion mechanism compared with simple splicing, and the calculation method of gain Δ_{attn} is the same as above.

The second set of comparisons addresses the model architecture innovations in section 3, comparing the performance of this paper to the standard Transformer baseline and other classical regression models. The standard Transformer baseline uses the original architecture described in Section 3.2 but does not contain the three innovations of Section 3.3 (transmission path prior bias, two-head attention decoupling, and frequency-time-domain union), and the hyperparameters remain consistent. Classical regression models include support vector regression (SVR, radial basis kernel is used for kernel function, and C and γ parameters are determined by grid search) and random forest regression (RF, number of trees $n_{tree} = 100$, maximum depth $d_{max} = 20$), both of which take 38-dimensional fusion features as input. In addition, in order to verify the advantages of deep learning over traditional machine learning, the psychoacoustic parameter linear regression model is added as the benchmark lower limit. The model only takes four psychoacoustic parameters as input to fit the linear regression equation:

$$\hat{y}_{lin} = \beta_0 + \beta_1 N + \beta_2 S + \beta_3 R + \beta_4 F \quad (70)$$

β_0 is the intercept term, and β_1 to β_4 are the regression coefficients of loudness N , sharpness S , roughness R and fluctuation intensity F , respectively, which are estimated by the least squares method. [Table 7](#) summarizes the configuration points and expected comparison conclusions of each comparison model.

Table 7. Comparison of model configurations in different experimental groups

Group of comparison	Model/Configuration	Feature of input	Difference in core
Ablation of modes	Acoustic only model	48-dimensional acoustic features	No vibration structure information
Ablation of modes	Vibration only model	54-dimensional vibration characteristics	Silent learning information
Ablation of modes	Simple splicing model	102-dimensional direct splicing	No attention fusion mechanism
Ablation of architecture	Standard Transformer	38-dimensional fusion features	No physical priors and two-headed design
The traditional approach	SVR	38-dimensional fusion features	Kernel methods, non-deep architectures
The traditional approach	RF	38-dimensional fusion features	Ensemble tree, non-sequential modeling
Lower base limit	Psychoacoustic linear regression	4-dimensional psychoacoustic parameters	No multimodal, no deep learning

Based on the two sets of comparative experiments, the modal ablation experiment is expected to reveal that the vibration mode has the highest contribution rate (supplementing the structural transmission information that acoustic features cannot reflect), and attention fusion brings an accuracy improvement of about 10% to 15% compared with simple splicing; the architecture ablation experiment is expected to verify that the combined contribution of the three innovative modules reduces MAE by about 20%. [Table 8](#) summarizes the quantitative results of each set of comparative experiments, where all MAE and MAPE are presented as the mean $\pm$ standard deviation of the three sets of repeated experiments to ensure statistical rigor.

Table 8. Expected output format of quantitative assessment indicators for each comparison experimental group

Group of comparison	Evaluate subitems	The loader MAE	The roller MAE	The excavator MAE
Ablation of modes	Acoustic only model	0.48 ± 0.05	0.52 ± 0.06	0.57 ± 0.06
Ablation of modes	Vibration only model	0.52 ± 0.05	0.55 ± 0.05	0.60 ± 0.07
Ablation of modes	Simple splicing model	0.38 ± 0.04	0.42 ± 0.05	0.47 ± 0.05
Ablation of architecture	Standard Transformer	0.39 ± 0.04	0.43 ± 0.04	0.48 ± 0.05
The traditional approach	SVR	0.56 ± 0.06	0.60 ± 0.07	0.65 ± 0.07
The traditional approach	RF	0.50 ± 0.05	0.54 ± 0.06	0.58 ± 0.06
Lower base limit	Psychoacoustic linear regression	0.72 ± 0.08	0.78 ± 0.09	0.82 ± 0.10
Complete method of this paper	Improved Transformer+ multimode	0.31 ± 0.03	0.36 ± 0.03	0.40 ± 0.04

The whole verification process covers the complete link from system integration to stand-alone verification and then to comparative demonstration. All test data, model weights and test codes have been archived and encapsulated to form a reproducible engineering verification benchmark. The prototype system runs stably on an industrial control computer equipped with Intel Core i7-12700H processor and 16GB memory. In real-time mode, the CPU occupancy is less than 35% and the memory occupancy is less than 2.5GB, which meets the resource constraints of engineering field deployment.

6.DISCUSSION

The multi-mode perception-evaluation-optimization integrated framework constructed in this paper shows good prediction accuracy and optimization effect in the experimental verification, but there are still some academic problems and engineering constraints worth further discussion after in-depth review of the whole technology chain. First of all, it is worth paying attention to the validity boundary of the physical prior embedding in the feature fusion layer. In this paper, the cosine similarity of the contribution rate of the transmission path is used as the attention bias term. Its core assumption is that the acoustic events dominated by the same transmission path have similar perception contributions in the timing, but this assumption faces challenges in the operation cycle. Because the main excitation sources of the loader in the two stages of shovel and unloading are the hydraulic system and the gearbox

respectively, the path contribution rate vector of the two is significantly different but the sound quality score may be similar. At this time, the constraint imposed by the attention bias term may limit the learning freedom of the model for the complex interaction mode across paths. The future improvement direction is to change the path similarity deviation into learnable dynamic weight rather than fixed cosine similarity, so that the model can independently judge when to rely on physical prior and when to rely on data driven. Secondly, the 200 groups of simulation samples used in the construction stage of the proxy model are still sparse in the 8-dimensional parameter space. Although the Gaussian process regression itself has a good adaptability to small samples, there is a significant interaction effect between the plate thickness and the stiffening height of the construction machinery structure. For example, increasing the thickness of the floor while reducing the height of reinforcement may produce acoustic radiation characteristics completely different from the traditional design experience. This high-order interaction is difficult to be fully captured by the GPR kernel function under the condition of finite samples, and the resulting uncertainty of the agent model may cover up the real Pareto front shape in the later stage of optimization search.

Another issue that needs careful discussion is the size and annotation consistency of the subjective evaluation database. The 73 valid samples retained after strict screening in this paper are statistically insufficient for training a Transformer network with more than two million parameters, although the data enhancement strategy alleviates the risk of overfitting to a certain extent. However, the enhanced samples are highly correlated in the feature space, which cannot really increase the information entropy of data distribution, which is also the fundamental reason for the difference between the verification set MAE in [Table 8](#) among the three models. The loader and the training set are of the same model, so the distribution overlap degree is high and the error is minimum, while the excavator has the largest structural difference and the error is relatively high. This finding suggests that in future engineering applications, at least 30 to 50 annotation samples of the new model should be collected for fine-tuning before the evaluation model is deployed for the new model, rather than directly using pre-training weights for zero-sample inference. In addition, although the two dimensions of annoyance and preference in subjective evaluation are highly correlated, they show a separation trend in some design schemes, that is, an optimization scheme may not improve preference while reducing annoyance. The weighted comprehensive improvement index in this paper may not reflect the differentiated auditory demands of different user groups by using fixed weights. There is a systematic difference between the preference of sound quality under operating conditions and the preference of sound quality under quality inspection conditions. Subsequent research can introduce Bayesian hierarchical models to conduct hierarchical modeling of evaluators with different operating experience.

From the perspective of engineering implementation, the optimization scheme in this paper is relatively easy to realize in terms of plate thickness adjustment and reinforcement layout. However, there is a critical relationship between the optimal value of mounting stiffness and the coupling mode frequency of engine-gearcase-cab. Adjusting mounting stiffness may introduce the resonance frequency originally avoided into a new speed range. This risk is covered to some extent in the optimization framework of this paper through the finite element simulation database, but the simplified treatment of the frequency varying stiffness and hysteresis damping characteristics of the rubber mount in the simulation model may lead to prediction bias. Finally, it is worth mentioning the universality boundary of the method. The loaders, rollers and excavators selected in this paper are all typical non-highway construction machinery, whose common features are low-speed and high-torque power system and rigid frame structure. For crawler cranes or concrete pump trucks with long hydraulic pipelines and rotary platforms, the dominant mechanism of sound quality may shift from structural radiation noise to fluid

noise and hydraulic pulsating noise. At this time, the feature set of vibration transfer path in this paper needs to supplement fluid characteristic dimensions such as hydraulic pipeline pulsating pressure and pump outlet flow fluctuation before it can be effectively transferred and applied.

7. CONCLUSION

Focusing on the evaluation, prediction and optimization design of sound quality of construction machinery products, this paper carries out systematic research from three levels of multi-mode perception feature fusion, in-depth evaluation model construction, sensitive parameter identification and structure optimization, and forms a complete theoretical method and engineering application technology system. First, at the feature level, an acoustical and vibration multi-mode feature extraction method covering psychoacoustic parameters, VMD modal energy, auditory spectrum coefficient, transmission path contribution rate and structural mode parameters is proposed. Through multi-head attention fusion and mRMR screening, the original 102-dimensional features are compressed to 38 dimensions, and the average redundancy is reduced to less than 0.12. The unification of physical completeness and computational efficiency is realized. Second, at the evaluation model level, aiming at the dilemma that deep networks are easy to overfit under the condition of small samples, an improved Transformer architecture embedding physical prior attention deviation of transmission path, decoupling of two-headed attention function and frequency-time-domain joint attention is proposed. Under the condition of training with only 73 labeled samples, the model can achieve the prediction accuracy of MAE of the validation set of 0.31-0.40, which is about 56% lower than that of the traditional psychoacoustic linear regression model and 20% lower than that of the standard Transformer baseline, and the increase of cross-model generalization error is controlled within 27%. It proves that physical knowledge embedding can substantially improve the generalization ability of deep models. Thirdly, at the optimization design level, a complete optimization chain of sensitive parameter location-GPR agent model acceleration-improvement NSGA-III multi-objective solution driven by SHAP is established. Three key frequency subbands (corresponding to the second and fourth order of the engine and the engagement frequency of the gearbox) and two dominant structural paths (the front and rear suspension paths of the cab) that affect the acoustic quality of the loader are successfully identified. The Pareto front solution provides nine effective non-dominated schemes under the constraint of mass incremental. The real transformation verification shows that the comprehensive improvement indexes of idle speed, uniform speed, acceleration and operation cycle are 0.52, 0.68, 0.91 and 0.73 respectively, and the maximum prediction deviation of the proxy model is 0.27, which verifies the reliability of the optimization method at the physical prototype level. Fourthly, through the system verification of three typical models of loader, roller and excavator, the general applicability of the proposed method in construction machinery products with significant differences in power transmission forms is proved, which provides a quantitative tool and methodological reference for the transformation of construction machinery industry from empirical trial and error acoustic improvement to data-driven forward design.

Abbreviations

mRMR, maximum correlation minimum redundancy;

SHAP, SHapley Additive exPlanations;

NSGA-III, Nondominated Sorting Genetic Algorithm III;

TPA, Transfer Path Analysis;
OTPA, Operational Transfer Path Analysis;
VMD, Variational Mode Decomposition;
MFCC, Mel-frequency Cepstral Coefficient;
GFCC, Gammatone Cepstral Coefficient;
PolyMAX, Polyreference Maximum Likelihood;
GPR, Gaussian Process Regression;
EI, Expected Improvement;
SBX, Simulated Binary Crossover;
SNR, Signal-to-Noise Ratio;
ICC, Intraclass Correlation Coefficient;
MSE, Mean Squared Error;
MAE, Mean Absolute Error;
MAPE, Mean Absolute Percentage Error;
RMSE, Root Mean Square Error;
SVR, Support Vector Regression;
RF, Random Forest;
ReLU, Rectified Linear Unit;
GELU, Gaussian Error Linear Unit;
L2, L2 Regularization;
CPU, Central Processing Unit;
GPU, Graphics Processing Unit;
NI, National Instruments;
ECU, Electronic Control Unit;
CAN, Controller Area Network;
ROPS, Roll-Over Protective Structure;
ARD, Automatic Relevance Determination;
HL, Hearing Level;
ITU-R, International Telecommunication Union - Radiocommunication Sector;
ISO, International Organization for Standardization;

GB/T, Guobiao/Tuijian (Chinese National Standard/Recommendation).

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **J.Z.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **J.Z.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

- [1] Peng, F., Li, W., Li, H., Yan, F., & Ye, J. (2025). A Subjective Sound Quality Evaluation Method for Vehicle Interior Noise Based on Feature Ranking and Minimizing Inter-Group Differences. *Results in Engineering*, 107643. <https://doi.org/10.1016/j.rineng.2025.107643>
- [2] Asady, H., Forouharmajd, F., Pourabdian, S., Jalali, M., Ahmadi, A., Ansari, F., & Fuente, A. (2025). Ear canal role and psychoacoustic parameters: loudness, sharpness, roughness, and fluctuation strength. *International Journal of Environmental Health Engineering*, 14(2), 10. https://doi.org/10.4103/ijehe.ijehe_11_24
- [3] Zhao, T., Chen, G., Suraphee, S., Phoophiwfa, T., & Busababodhin, P. (2025). A hybrid TCN-XGBoost model for agricultural product market price forecasting. *PLoS One*, 20(5), e0322496. <https://doi.org/10.1371/journal.pone.0322496>
- [4] Hong, J. Y., Lam, B., Ong, Z. T., Ooi, K., Gan, W. S., & Lee, S. (2022). A multidimensional assessment of construction machinery noises based on perceptual attributes and psychoacoustic parameters. *Automation in Construction*, 140, 104295. <https://doi.org/10.1016/j.autcon.2022.104295>
- [5] Zhao, T., Chen, G., Pang, C., & Busababodhin, P. (2025). Application and performance optimization of SLHS-TCN-XGBoost model in power demand forecasting. *Comput. Model. Eng. Sci*, 143(3), 2883-2917. <https://doi.org/10.32604/cmesci.2025.066442>
- [6] Shao, H., Lin, J., Zhang, L., Galar, D., & Kumar, U. (2021). A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance. *Information Fusion*, 74, 65-76. <https://doi.org/10.1016/j.inffus.2021.03.008>
- [7] Zhao, T., Chen, G., Pang, C., Seenoi, P., Papukdee, N., & Busababodhin, P. (2025). Time-lapse earthquake difference prediction based on physics-informed long short-term memory coupled with interpretability boosting. *Journal of Seismic Exploration*, 34(3), 25. <https://doi.org/10.36922/JSE025310049>
- [8] Zhao, T., Chen, G., Pang, C., Seenoi, P., Papukdee, N., Busababodhin, P., & Du, Y. (2025). Hybrid convolutional neural network-graph attention network-gradient boosting decision tree model for seismic impedance inversion prediction. *J Seismic Explor*, 34(5), 81-

98. <https://doi.org/10.36922/JSE025310051>

[9] Zhang, Y., Tian, J., & Xiong, Q. (2025). A review of embodied intelligence systems: A three-layer framework integrating multimodal perception, world modeling, and structured strategies. *Frontiers in Robotics and AI*, 12, 1668910. <https://doi.org/10.3389/frobt.2025.1668910>

[10] Zhao, T., Chen, G., Pang, C., Li, L., & Busababodhin, P. (2026). Forecasting global agricultural trade imbalances using a hybrid deep learning and gradient boosting framework. *Discover Computing*, 29(1), 443. <https://doi.org/10.1007/s10791-026-10367-8>

[11] Zhao, T., Chen, G., & Busababodhin, P. (2026). An intelligent expert system for logistics disruption prediction and mitigation in global supply chains: Integrating graph-based risk inference with ensemble forecasting. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-61617-0>

[12] Bosia, F., Dal Poggetto, V. F., Gliozzi, A. S., Greco, G., Lott, M., Miniaci, M., ... & Pugno, N. M. (2022). Optimized structures for vibration attenuation and sound control in nature: A review. *Matter*, 5(10), 3311-3340. <https://doi.org/10.1016/j.matt.2022.07.023>

[13] Zhao, T., Chen, G., Li, J., Pang, C., Seenoi, P., Papukdee, N., & Busababodhin, P. (2026). XGBoost-Deep Residual (FedXGB-ResNet) collaborative porosity prediction in a federated learning framework. *Journal of Seismic Exploration*, 35(2), 201. <https://doi.org/10.36922/JSE025440094>

[14] Xie, W., & Hua, Y. (2024). Structural vibration comfort: A review of recent developments. *Buildings*, 14(6), 1592. <https://doi.org/10.3390/buildings14061592>

[15] Du, Y., Chen, G., Pang, C., & Zhao, T. (2025). Prediction of fracture and vug parameters in carbonate reservoirs using a combined T-GNO-PINN approach. *Journal of Seismic Exploration*, 35(1), 46. <https://doi.org/10.36922/JSE025330057>

[16] Zhang, Z., Jia, Z., Wei, J., Xiao, Y., Zhang, R., & Zhao, T. (2025). Combining 5G Communication Technology with XGboost Model for Global Gas Trade Flow Prediction. In *International Conference on Artificial Intelligence and Communication Technology* (pp. 53-64). Springer. https://doi.org/10.1007/978-3-032-06376-2_5

[17] Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *Ieee Access*, 10, 122136-122158. <https://doi.org/10.1109/ACCESS.2022.3223444>

[18] Bezoui, M., Elmoutaouakkil, A., & Beni-hssane, A. (2016, September). Feature extraction of some Quranic recitation using mel-frequency cepstral coefficients (MFCC). In *2016 5th international conference on multimedia computing and systems (ICMCS)* (pp. 127-131). IEEE. <https://doi.org/10.1109/ICMCS.2016.7905619>

[19] Ma, L., Jiang, A., & Jiang, W. (2024). The intelligent diagnosis of a hydraulic plunger pump based on the MIGLCC-DLSTM method using sound signals. *Machines*, 12(12), 869. <https://doi.org/10.3390/machines12120869>

[20] Wang, Z., Chen, Y., Wang, H., Wang, H., & Fang, F. (2024, August). Acoustic feature extraction of wind turbine blades based on improved mfcc-gfcc. In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)* (pp. 1-6). IEEE. <https://doi.org/10.1109/INDIN58382.2024.10774206>

[21] Shang, Z., Hu, F., Zeng, F., Wei, L., Xu, Q., & Wang, J. (2021). Research of transfer path analysis based on contribution factor of sound quality. *Applied acoustics*, 173, 107693.

<https://doi.org/10.1016/j.apacoust.2020.107693>

- [22] Liu, N., Sun, Y., Wang, Y., Sun, P., Li, W., & Guo, H. (2021). Mechanism of interior noise generation in high-speed vehicle based on anti-noise operational transfer path analysis. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 235(1), 273-287. <https://doi.org/10.1177/0954407020937219>
- [23] Khakshournia, S., Haghighi, S. S., Majidi, M., Najafnia, F., & Khodaparast, H. H. (2026). A hybrid strategy of OTPA and machine learning for efficient root-cause analysis of NVH in multi-source systems. *Applied Acoustics*, 242, 111060. <https://doi.org/10.1016/j.apacoust.2025.111060>
- [24] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8), 1226-1238. <https://doi.org/10.1109/TPAMI.2005.159>
- [25] Wang, X., Tao, Y., & Zheng, K. (2018, July). Feature selection methods in the framework of mRMR. In *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)* (pp. 1490-1495). IEEE. <https://doi.org/10.1109/IMCCC.2018.00307>
- [26] Qu, K., Xu, J., Han, Z., & Xu, S. (2023). Maximum relevance minimum redundancy-based feature selection using rough mutual information in adaptive neighborhood rough sets. *Applied Intelligence*, 53(14), 17727-17746. <https://doi.org/10.1007/s10489-022-04398-z>
- [27] Man, X., Wan, Y., Han, B., Wei, Y., Kong, L., Zhang, Z., & Wang, J. (2025). Wavelet and time domain attention-guided stochastic resonance network for fault diagnosis of rotating machinery in strong noise environment. *Measurement Science and Technology*, 36(7), 076122. <https://doi.org/10.1088/1361-6501/adeceb6>
- [28] Gong, T., Liu, Z., Xie, C., & Yang, Z. (2019). Property investigation on the additive white Gaussian noise after sub-Nyquist sampling. *IEEE Access*, 7, 122820-122826. <https://doi.org/10.1109/ACCESS.2019.2938023>
- [29] Yuan, Z., Meng, L., Gu, X., Bai, Y., Cui, H., & Jiang, C. (2021). Prediction of NO_x emissions for coal-fired power plants with stacked-generalization ensemble method. *Fuel*, 289, 119748. <https://doi.org/10.1016/j.fuel.2020.119748>
- [30] Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, 127, 1-10. <https://doi.org/10.1016/j.apenergy.2014.04.016>

APA Citation

Zhou, J. (2026). Acoustic Quality Evaluation Model and Optimization Design of Construction Machinery Products Based on Multi Modal Perception. *Journal of Discovery Core*, 1(1), 163-199. <https://doi.org/10.67541/jdc2607>