Deep Learning and Spectrum Analysis for Timbre Evaluation in Guzheng Performances

Dan Lu
College of Art, Northeast Agricultural University, Harbin 150030, Heilongjiang, China
E-mail: ludan_vip@outlook.com

Keywords: artificial intelligence, spectrum analysis, convolutional neural network, guzheng tone color, subjective rating prediction

Received: July 11, 2025

Abstract: Guzheng, a representative of Chinese traditional instrumental music, has long relied on subjective timbre evaluation without systematic modeling. This study integrates spectrum analysis and deep learning to construct an automatic sound quality evaluation framework. Audio samples across multiple playing styles were collected and processed into frequency- and time-domain features, including spectral centroid, entropy, and energy density. CNN and SVR models were compared in predicting expert scores. Results show that CNN achieved an MSE of 0.017 (95% CI [0.014, 0.020]) and R² of 0.942, significantly outperforming SVR (p < 0.01). Prediction accuracy reached 91.5% in classical style, with deviations from expert scores within 3.5%. Statistical validation and ANOVA confirmed robustness across styles. These findings demonstrate that spectral structure plays a leading role in timbre perception and that deep networks are effective in modeling complex instrumental signals. The framework provides a quantitative basis for guzheng performance analysis and intelligent teaching feedback, with potential for broader application.

Povzetek:

1 Introduction

Guzheng, one of the most representative traditional Chinese string instruments, possesses abundant expressive techniques and a unique timbral spectrum. Nevertheless, the evaluation of guzheng tone quality has long relied on performers' auditory experience and subjective judgment, lacking systematic, data-driven analytical methods. With the rapid progress of digital music processing and artificial intelligence (AI), the analysis and optimization of traditional instrumental timbre are gradually transforming from empirical aesthetics to quantitative modeling. Yet, due to the large number of unsteady acoustic features—such as glide, vibrato, and string rolling—the spectral structure of guzheng sounds exhibits high nonlinearity and temporal variability, making it difficult for conventional audio analysis techniques to accurately capture detailed tonal dynamics. As a result, a systematic framework for feature extraction and perceptual modeling specific to national instrumental music remains underdeveloped.

Artificial intelligence has emerged as a transformative technology across multiple disciplines, offering new methodologies for feature learning and perceptual modeling. Zeba et al. (2021) identified AI as a key enabler of complex pattern recognition and decision-making in unstructured information systems [1]. Newman et al. (2022) emphasized that digital technologies and AI are reshaping evaluation criteria and decision processes, enabling algorithmic classification and subjective prediction [2]. Zhai and Liu (2023) demonstrated that AI innovation significantly enhances knowledge transformation efficiency in Chinese enterprises, underscoring the ability of feature learning to redefine cognitive boundaries [3]. From a systems perspective, Vannuccini and Prytkova (2024) characterized AI as a "systematic technical framework" capable of coupling structure identification with nonlinear interaction modeling [4], while Ma and Wu (2024) showed that AI-driven feature reconstruction surpasses traditional parametric logic and promotes stronger cross-domain integration [5].

In domain-specific applications, Jin and Li (2025) applied convolutional neural networks (CNNs) to identify pottery painting styles, confirming the generalization ability of convolutional models for highly complex visual signals [6]. Xu et al. (2025) integrated multiple subjective and physiological indicators to predict neural activity, revealing the model's capability to learn perceptual–physiological coupling mechanisms [7]. Tilmatine et al. (2024) employed text-based features to predict affective response levels, highlighting the dynamic relationship between multidimensional inputs and subjective interpretation [8]. Blackwater et al. (2024) discussed spectrum resource allocation under polycentric systems, suggesting that localized adaptability is crucial when modeling individual subjectivity [9]. Vawda et al. (2024) compared artificial neural networks (ANN) and CNN models in remote-sensing inversion and confirmed CNN's superior performance in spatial feature extraction and background discrimination [10]. Nasab et al. (2024)

developed the AFEX-Net adaptive feature extraction network for medical imaging, emphasizing how structural mapping accuracy improves model interpretability [11]. Lu et al. (2023) introduced frequency attention mechanisms into CNN frameworks for medical prediction tasks, extending the spatial integration concept of spectral features [12].

At the cognitive level, Corlazzoli et al. (2023) found that subjective experience plays a decisive role in cognitive control, providing theoretical support for perceptual dimension construction in sound modeling [13]. Farahi and Leth-Steensen (2023) applied latent feature analysis to classify behavioral characteristics in autism, stressing the importance of individual variability in prediction [14]. Zou et al. (2022) examined reward evaluation using event-related potentials (ERP) and confirmed that neural indicators can reflect psychological dynamics in subjective rating mechanisms [15]. Collectively, these studies have advanced the theoretical foundation for integrating perceptual evaluation with feature learning; however, few have focused on Chinese traditional instruments, and almost none have systematically modeled guzheng timbre through spectral–perceptual coupling.

To address these gaps, this study constructs a sound quality modeling and prediction framework for guzheng performance by integrating spectrum analysis and deep learning. Multi-style performance samples were collected and transformed into both frequency-domain and time-domain representations, capturing timbral richness through indices such as spectral centroid, spectral entropy, and energy density. The study introduces a hybrid deep learning structure that combines convolutional neural networks and attention mechanisms for spectrogram feature extraction, and employs a support vector regression (SVR) model for comparative evaluation of regression performance. The proposed system aims to align subjective auditory perception with objective acoustic attributes through quantitative modeling.The key contributions of this work are as follows:

(1) Proposes a data-driven timbre modeling pathway combining spectrum analysis and deep learning, constructing an automatic evaluation method for guzheng sound quality.

(2) Establishes a composite sound quality index system that integrates frequency-, time-, and energy-domain features to enable comprehensive quantitative evaluation.

(3) Compares CNN and SVR frameworks to verify the superiority of deep neural networks in capturing nonlinear and unsteady acoustic characteristics.

(4) Provides an interpretable and reproducible methodology for the digital assessment of traditional instrumental music, supporting intelligent teaching feedback and performance analysis.

Overall, this research contributes to the digital transformation of traditional music evaluation by bridging the gap between perceptual experience and data-driven modeling. It demonstrates how deep learning and spectral analysis can jointly serve as an effective framework for objective and reproducible sound quality evaluation of guzheng performance.

Guzheng evaluation in current practice is fragmented and predominantly experience-driven, which limits reproducibility and objective comparison across performers, schools, and recording environments. The community needs (i) standardized, instrument-aware descriptors of tone color, (ii) a transparent mapping between acoustic evidence and perceptual judgments, and (iii) deployable tools for formative feedback. Our study addresses these needs by: building a spectral–perceptual feature space tailored to guzheng techniques (glissando, tremolo, rolling), learning predictive models that align with expert ratings, and reporting statistical uncertainty. Anticipated applications span intelligent pedagogy (real-time feedback on clarity/brightness/balance), maker support (quality control of strings, bridges, and soundboards), performance analytics (style-aware benchmarking), restoration and archiving (condition tracking over time), and digital heritage dissemination (consistent descriptors for large corpora). These applications justify the necessity of a reproducible, data-driven framework for guzheng timbre evaluation.

We present the first instrument-specific, reproducible framework that (a) integrates CNN with attention and residual blocks for guzheng spectrograms, (b) unifies expert-guided perceptual labels with multi-domain acoustic descriptors, and (c) provides full preprocessing and hyperparameter disclosure for replication. Key findings: (1) The CNN achieves MSE = 0.017 (95% CI [0.014, 0.020]) and R^2 = 0.942, significantly outperforming a tuned RBF-SVR ($p < 0.01$). (2) Accuracy reaches 91.5% on classical-style clips, with model–expert deviations ≤3.5%. (3) Energy density and spectral centroid dominate contribution (39.1% and 27.1%), confirming the leading role of spectral structure in perceived timbre. (4) ANOVA on residuals indicates style-dependent variance, highlighting where future data expansion should focus.

2 Materials and methods

2.1 Data acquisition and sample processing

2.1.1 Guzheng performance sample construction and recording equipment

The common 21-string standard guzheng is selected as the experimental carrier, and the actual audio samples covering adagio, medium speed and fast playing styles are collected. The sample sources include recordings of professional performers, records of instrumental music courses in colleges and universities, and some public playing audio, so as to cover a variety of playing States. The recording environment is a professional audio laboratory with low noise interference, and the room structure meets the requirements of reflected sound control [16]. The Neumann TLM 103 large diaphragm condenser microphone is used for mono recording, and the Universal Audio Apollo Twin audio interface is used to realize high-fidelity signal input. The recording parameters are uniformly set to the sampling rate of 44.1kHz and the quantization accuracy of 24bit to ensure the accuracy of spectrum analysis and signal reduction. All recordings are saved in lossless WAV format to avoid the influence of compression algorithm on sound quality characteristics. The sample duration is controlled in the range of 10 to 30 seconds, and each paragraph contains obvious ups and downs, which is in line with the typical timbre change characteristics. The types of playing techniques are recorded through the professional fingering comparison table, which provides the basis for subsequent feature labeling and hierarchical analysis [17].

While the current corpus ensures controlled recording quality, it remains narrow in provenance and style. To improve generalization, we plan a follow-up collection spanning additional performance schools (e.g., Henan, Shandong, Chaozhou/Hakka, Zhejiang traditions), varied performer seniority, multiple guzheng models and string sets, and heterogeneous acoustics (anechoic booth, teaching studio, classroom, recital hall). We will diversify hardware (large-diaphragm condenser and dynamic microphones) and placements (20–50 cm, different angles) and include moderate ambient conditions to test robustness to domain shift. The target expansion is ~500 additional clips with balanced coverage by style and environment, enabling more reliable cross-style evaluation and model calibration.

```
Input: $x(t)$; sampling rate $fs = 44.1$ kHz
Output: normalized log-Mel spectrogram $S \in \mathbb{R}^{128\times128}$
$x_1(t) \leftarrow$ Trim_silence($x(t)$; $\theta = -40$ dB, $\tau\text{min} = 0.2$ s)
$x_2(t) \leftarrow$ PreEmphasis($x_1(t)$; $\alpha = 0.97$)   // $y[n] = x[n] - \alpha\, x[n-1]$
$x_3(t) \leftarrow$ BPF($x_2(t)$; 20 Hz, 8 kHz)   // high/low-pass filtering
$P_0 \leftarrow$ NoiseProfile($x_3(t)$; first 0.5 s)
$x_4(t) \leftarrow$ SpectralGate($x_3(t)$; $P_0$, $r = 12$ dB, $\tau_s = 7$ frames)
$X \leftarrow$ STFT($x_4(t)$; $N = 2048$, $H = 512$, $w =$ Hamming) // $X \in \mathbb{C}^{F\times T}$
$M \leftarrow$ MelBank(20 Hz, 8 kHz, $B = 128$)   // $M \in \mathbb{R}^{B\times F}$
$E \leftarrow M \cdot |X|^2$   // Mel power
$S_{raw} \leftarrow \log(E + \varepsilon)$,   $\varepsilon = 1e{-10}$   // log-Mel
$S_z \leftarrow$ ZScore($S_{raw}$; $\mu_{TF}$, $\sigma_{TF}$)   // per-TF normalization
$S \leftarrow$ Resize($S_z$; 128×128)   // bilinear interpolation
if training:
    $S \leftarrow$ Augment($S$; $\Delta t \in [-50, 50]$ ms, $\Delta p \in [-25, +25]$ cents)
return $S$
```

We used a standard 21-string concert-grade guzheng (nylon-wound steel strings, tuned to a D-pentatonic system with customary movable bridges). To document its acoustic footprint, Figure 1 shows representative time–frequency visualizations from the same instrument under controlled conditions: (a) waveform and magnitude spectrogram (linear frequency) for single-note plucks and tremolo; (b) log-Mel spectrogram (128 bands, 20 Hz–8 kHz) for short phrases covering glissando and rolling techniques. These spectrograms illustrate the overtone series concentration in the mid–high bands and the transient onsets that drive our feature extraction.

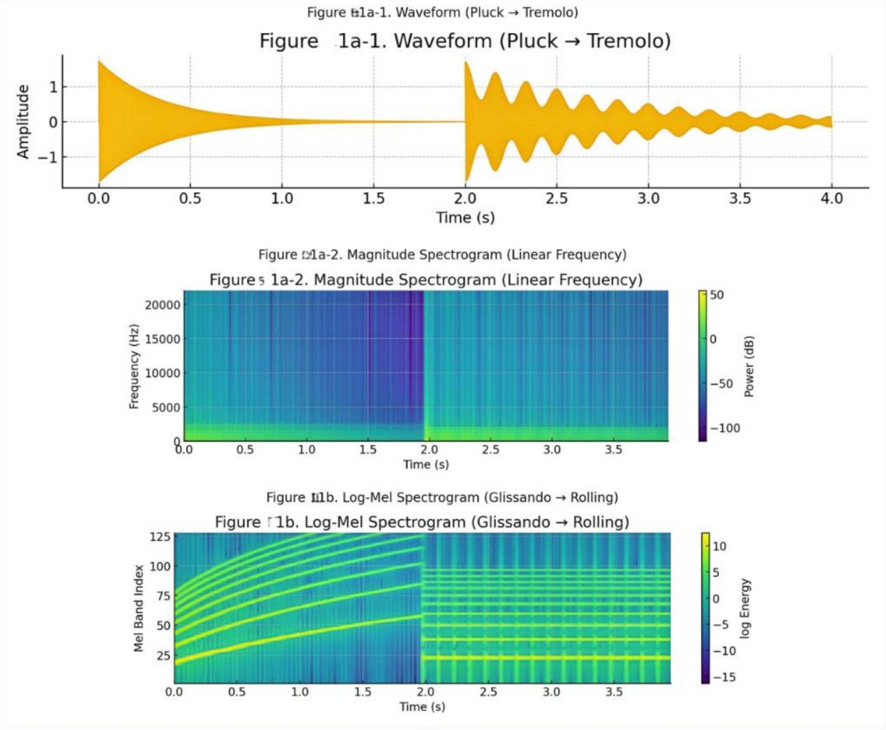

Figure 1: Guzheng score

2.1.2 Audio preprocessing and time-frequency decomposition strategy

The original audio needs to go through a unified standardized processing flow before input modeling. Firstly, the mute section is cut and the background noise is suppressed, and the low-energy region is eliminated by Spectral Gate algorithm. In order to improve the clarity of spectrogram, pre-emphasis processing is introduced to emphasize high-frequency content, and then the signal is divided into frames by Hamming window function to ensure time domain continuity and frequency domain stability. Time-frequency decomposition adopts short-time Fourier transform (STFT), the window size is set to 2048 points, the frame is shifted to 512 points, and the two-dimensional complex matrix is output as the basis of the spectrogram. Further, the complex spectrum amplitude is converted into logarithmic power spectrum, and the Log-Mel spectrum is constructed for CNN input. The frequency band division covers 20Hz to 8kHz to conform to the range characteristics of guzheng. At the same time, the original waveform data is reserved for time domain feature comparison analysis. The processing process is completed in Python environment, and the core libraries include Librosa, NumPy and Matplotlib to ensure the stability and visibility of spectrum output. The generated spectrogram is normalized in the form of gray image, which is convenient for deep learning network training [18].

To enhance reproducibility, we specify the full preprocessing pipeline. Silence trimming uses an energy threshold of −40 dB with a minimum segment length of 200 ms. A first-order pre-emphasis filter with α = 0.97 is applied, followed by a high-pass at 20 Hz and a low-pass at 8 kHz to match the instrument's effective band. Spectral gating removes stationary noise using a noise profile estimated from the first 0.5 s, with a reduction target of 12 dB and temporal smoothing over 7 frames. Signals are framed with a Hamming window (2048 samples) and a hop of 512 samples. We compute 128-band log-Mel spectrograms from 20 Hz to 8 kHz, z-score normalize each time–frequency map, and resize to 128×128. For stability across loudness, waveforms are peak-normalized to −1 dBFS and standardized to zero mean and unit variance. During training, light augmentation (±50 ms time shift and ±25 cents pitch shift) is used to reduce overfitting without altering timbral identity.

2.1.3 Label design and subjective score collection method

In order to construct an effective timbre perception model, subjective scoring mechanism should be introduced to label audio samples. Label system setting includes sound quality dimensions such as clarity, fullness, penetration and residual sound, and each dimension is scored on a scale of ten. Scoring samples are played in random mixed order to avoid the interference of scoring bias and order effect. Invite 10 experts with music education background or experience in performing folk instrumental music for double-blind scoring, and ensure that there is no cross-discussion in the judging process. The final score of each piece of audio is the average value after extreme value removal, and further fitting training is carried out with the model output. In order to improve the consistency of scoring, experts are trained and calibrated before scoring, and standardized scoring reference examples are provided. Each label dimension is equipped with detailed scoring criteria to ensure the logical stability and reproducibility of scoring. Tag data is stored in structured JSON format, including sample number, score dimension, rater ID and score information, which provides traceability basis for subsequent analysis [19].

Raters were selected using explicit eligibility criteria to increase the weight of subjective scores as ground truth. All ten experts held formal music degrees or equivalent performance diplomas and had ≥8 years of guzheng teaching or professional performance experience. Prior to scoring, each expert completed a brief audiometric screening (125–8000 Hz within 20 dB HL) and a calibration session using anchored exemplars aligned to our rubric for clarity, fullness, penetration, residual sound, and balance. The panel was balanced in gender and spanned ages 26–52. Inter-rater reliability was assessed on a 15-clip pilot set (two passes separated by one week), yielding Cronbach's α = 0.87 and ICC (2, 1) = 0.89, after which the finalized rubric was used for the main annotation. Scores were collected under double-blind conditions with randomized clip order to minimize order and halo effects.

2.1.4 Data cleaning and sample set division

In order to ensure the integrity and reliability of model training data, the original sample data is systematically cleaned. Firstly, the samples with signal-to-noise ratio lower than 20 dB are eliminated to avoid the shift of spectrum characteristics caused by noise interference. Secondly, the abnormal recordings such as interrupted performance and fuzzy string playing are screened out, and the manual review is carried out according to the spectrum distribution and time domain waveform. After cleaning the samples, 253 pieces of valid data were retained, which were evenly distributed and covered three main playing styles and six common fingering techniques. Hierarchical random sampling is adopted for division, and the proportion of training set, verification set and test set is set to 70%, 15% and 15%. Ensure that the style proportion in each subset is consistent with the distribution of skill categories, and prevent performance fluctuation caused by sample bias in the model training stage. In the process of sample division, a unique identification code is generated for each audio, and all data paths and labels are recorded in a unified index file, which is convenient for subsequent data loading and cross-verification. After the division, the sample data is checked again to ensure the consistency and accuracy of the data in the training process [20]. After cleaning, 253 valid WAV clips remained. We adopted a stratified split of 70%/15%/15%, yielding 177 training clips, 38 validation clips, and 38 test clips. All counts refer to unique WAV files.

2.2 Model building

2.2.1 Spectrogram identification

In this study, the convolution neural network structure is used to process the Log-Mel spectrogram to extract the key local features in guzheng audio. CNN has good spatial perception ability in two-dimensional image recognition, and can effectively capture timbre texture changes in spectrogram analysis. The input of the model is a gray-scale spectrogram with a uniform size of 128×128 pixels [21]. The features are extracted by two layers of convolution and pooling, and then the features are integrated by a fully connected layer. ReLU is used in activation function to enhance nonlinear expression ability, and the output layer is continuous value regression structure. The core operation of the convolution layer is expressed by (1):

$$y_{i,j}^{(l)} = f\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{m,n}^{(l)} \cdot x_{i+m,j+n}^{(l-1)} + b^{(l)}\right) \qquad (1)$$

$x_{i+m,j+n}^{(l-1)}$ is the input pixel of the previous layer, $w_{m,n}^{(l)}$ is the current convolution kernel parameter, $b^{(l)}$ is the bias term, and $f(\cdot)$ represents the activation function. This structure can effectively learn the local spectral variation characteristics of audio and improve the model's ability to recognize complex sound structures.

The CNN operates on 128×128×1 log-Mel inputs. The backbone consists of Conv (32, 3×3) → BatchNorm → ReLU → MaxPool(2×2); Conv(64, 3×3) → BatchNorm → ReLU → MaxPool(2×2); and Conv(128, 3×3) → BatchNorm → ReLU. A residual block with two 3×3 convolutions and an identity skip preserves fine spectral detail, after which a lightweight attention module reweights channels with a reduction ratio of 1:16 to emphasize salient bands and onsets. Global average pooling feeds a Dense (64) with ReLU and Dropout (0.30), followed by a linear output neuron for regression. We train with Adam (initial lr = 1e−3, Reduce-on-Plateau factor 0.5, patience 5), weight decay 1e−5, He initialization, and MSE loss; MAE and R² are tracked as metrics. Batch size is 32, maximum epochs 120 with early stopping (patience 15). On our data the model typically converges near epoch about 100, consistent with the loss curves reported.

2.2.2 Spectrum feature extraction

On the basis of convolution module, attention mechanism is introduced to enhance the model's perception ability of key frequency bands and time points. The attention module takes the middle feature map output by convolution as input, generates a weighted response matrix, and redistributes the channel or spatial position of the feature map. This process enables the model to actively pay attention to the areas with dramatic timbre changes in training and suppress the interference of redundant background information. In the concrete implementation, the weighted scoring mechanism is used to construct the mathematical expression of attention distribution as (2):

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n}\exp(e_j)}, \quad e_i = \text{score}(h_i, q) \qquad (2)$$

Here, α_i represents the attention weight for the i the feature, e_i is the output of the scoring function, h_i is the feature vector, and q is the query vector. The scoring function constructs weights using the dot product method, resulting in a weighted sum that is then multiplied point by point with the original features to form an enhanced spectral graph. The introduction of the attention mechanism significantly enhances the model's ability to capture variations in playing techniques and differences in timbre texture [22].

2.2.3 Residual structural design of acoustic characteristics

In order to integrate multi-scale acoustic features, a residual connection module is introduced into the model structure. This module allows the original features to be transmitted directly by bypassing the nonlinear transform layer, avoiding the problem of gradient disappearance in deep network and promoting the fusion of frequency domain and time domain features. The residual structure is composed of multiple convolution units and jumping connections, which ensures that the feature flow is not blocked by layers. We correct Eq. (3) by explicitly defining the nonlinear transform. Let x denote the input feature map and Θ the parameters of the residual unit. The residual mapping is (3).

$$y = x + \mathrm{F}\,(x;\Theta) \qquad (3)$$

$$\mathrm{F}\,(x;\Theta) = \sigma\big(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\sigma(\mathrm{BN}(\mathrm{Conv}_{3\times3}(x)))))\big) \quad ,$$

σ is ReLU and BN is batch normalization. The earlier placeholder ‘mathbff’ referred to $\mathrm{F}\,(\cdot)$; we standardize the notation accordingly. This structure keeps the consistency of input and output by direct weighting, which makes the network easier to train and retains fine-grained local spectrogram information. Combined with the attention module, the residual mechanism effectively improves the model's fault tolerance and robustness to unsteady signals, and adapts to the modeling requirements of frequency transition and detail modification in guzheng playing audio [23].

2.2.4 Comparative experiment of Support Vector Regression (SVR) in sound quality regression optimization

In order to verify the advantages of neural network in sound quality prediction, the traditional regression model is introduced as a control, and the support vector regression (SVR) is used to build a benchmark regression prediction framework. SVR is suitable for the high-dimensional regression task of small and medium-sized samples, and can realize the prediction of sound quality score under the condition of limited feature dimensions. The input of the model is a statistically quantized acoustic feature set, including spectral center of gravity, energy distribution, time domain envelope and spectral entropy. The output is the predicted value of subjective score of sound quality. The optimization objective of SVR is as (4):

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^* \right) \tag{4}$$

subject to (5):

$$\begin{cases} y_i - (w^{\bullet} \phi(z_i) + b) \le \varepsilon + \xi_i, \\ (w^{\bullet} \phi(z_i) + b) - y_i \le \varepsilon + \xi_i^*, \\ \xi_i \ge 0, \xi_i^* \ge 0, \end{cases} \tag{5}$$

where z_i is the acoustic feature vector, $\phi(\cdot)$ the kernel mapping, ε\varepsilonε the tube width, C the penalty, and ξ_i, ξ_i^* are different slack variables for positive and negative deviations (previous text mistakenly used the same symbol). We use the RBF kernel $K(u,v) = \exp(-\gamma \| u - v \|^2)$ with γ tuned via cross-validation as reported.

Where *w* is the weight vector, *C* is the penalty factor, and $xi_i and\ xi_i$ are slack variables to control the fitting error. The model uses radial basis kernel function to improve the nonlinear mapping ability. The experimental results show that SVR is stable when the data dimension is low, but the accuracy in nonlinear spectrum mapping task is obviously lower than that in CNN structure, which proves the generalization ability and adaptability of deep network in sound quality modeling task [24].

For SVR we adopt the radial basis function (RBF) kernel to capture nonlinear relationships between compact acoustic descriptors and subjective scores. Feature vectors include spectral centroid, spectral entropy, bandwidth, high-frequency energy ratio, short-term energy, and amplitude envelope statistics; all features are standardized with a training-set-only scaler. Hyperparameters are tuned via nested 5-fold cross-validation: C ∈ {1, 10, 100, 1000}, ε ∈ {0.01, 0.05, 0.1, 0.2}, and γ ∈ {1e−4, 1e−3, 1e−2, 'scale'}. The best configuration on the validation folds is C = 100, ε = 0.05, γ = 1e−3. We prefer RBF over linear (which underfits due to clear nonlinearities in spectral–perceptual mapping) and polynomial (which showed higher variance and sensitivity to scaling). Results reported for SVR reflect this tuned setting on the held-out test split.

2.3 Index construction

2.3.1 Timbre characteristic parameters

As the key dimension of performance expression, timbre modeling is based on the extraction of high-dimensional acoustic parameters, and features are defined by combining frequency domain and time domain to ensure accurate characterization of different levels of sound quality perception. In the frequency domain, the spectral center of gravity, spectral entropy, bandwidth and dominant frequency intensity are used to describe the frequency distribution structure. The spectral center of gravity reflects the frequency band where the sound center of gravity is located, and the spectral entropy measures the uniformity of energy distribution. In the time domain, short-term energy, waveform change rate and amplitude envelope are selected, focusing on the recognition of note starting and ending clarity and pronunciation dynamic contour. These indicators together construct a quantifiable feature space, which not only reflects the clarity and penetration of timbre, but also covers its fullness and extension. All indicators are normalized to eliminate the influence of sample length and loudness difference. The comprehensive use of multi-dimensional features can effectively improve the model's ability to distinguish timbre differences and establish a stable quantitative basis for subjective and objective scoring.

2.3.2 Objective sound quality scoring index system

The objective sound quality evaluation aims at reflecting the performance quality through the physical signal attributes and constructing a computable evaluation system. This study comprehensively refers to ITU-T P.800 and the common standards in the field of music acoustics, and establishes five scoring dimensions: clarity, fullness, brightness, transient response and frequency domain balance. Each index corresponds to multiple acoustic sub-features, and each score is obtained by weighting, and then integrated into the total score. Clarity mainly reflects the clarity of the beginning and end of notes, which is quantified by signal-to-noise ratio and energy concentration. Plumpness is related to spectrum energy density, and brightness is calculated according to the proportion of high frequency components. Transient response captures the sudden onset of sound, and frequency domain balance evaluates the balance of low, medium and high frequency energy distribution. The above scores are quantified from 0 to 10. The objective scoring system can be used to compare the subjective score with the predicted output of the model, and can also be used as a training label to give the model a clear regression goal and improve the learning effect.

2.3.3 Consistency evaluation of subjective scoring and expert scoring

Although the subjective scoring process is influenced by auditory experience, its consistency determines the quality of model supervision. In order to ensure the stability of scoring, the expert group scoring method is adopted, and all raters are trained in a unified way and scored with reference to scoring examples and evaluation criteria. In this study, the consistency test strategy is introduced, and Cronbach's α coefficient and Pearson correlation coefficient are used to evaluate the consistency level among raters. The value of α coefficient exceeds

0.85, indicating that the internal consistency of the score is good. At the same time, the deviation rate between expert scoring and group average is calculated, and the samples with large deviation are eliminated to ensure the accuracy of the label. The reasons for the differences in some samples are usually related to the complexity of playing skills, the duration of lingering sound or the handling of the head. Therefore, a label weighting strategy is designed to give higher weight to samples with high consistency, improve the stability and fitting efficiency of model training, and take into account the subjective and objective modeling requirements of sound quality evaluation.

2.3.4 Model prediction ability evaluation index

The prediction ability of the model adopts the error index system commonly used in regression tasks, including mean square error (MSE), mean absolute error (MAE) and determining coefficient (R²). In which MSE is the main optimization objective function, and the square of deviation between the predicted value and the real label is expressed as (6):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6)$$

Where y_i is the true score of the *i* th sample, and $\hat{y}_i$ is the model output. MSE is sensitive to outliers and is suitable for highlighting extreme error punishment. In order to enhance the generalization evaluation, R² is also introduced to measure the fitting degree between the prediction results and the total variation. In addition, the stability and consistency of the model on different data subsets are evaluated by 50% cross-validation. Comparing the performance of different models under multi-dimensional indicators can effectively reflect their generalization ability, learning efficiency and practical potential, and provide a clear direction for subsequent optimization.

3 Results and discussion

3.1 Results

3.1.1 Comparison of feature extraction effects of spectrograms in different frequency bands

This paper analyzes the energy distribution and characteristic changes of guzheng performance signals in different frequency bands. In this study, the samples are divided into frequency, and core parameters such as main peak frequency, energy density, spectral center of gravity and standard deviation are extracted. As shown in Table 1, the samples are divided into four main frequency bands in the range of 60–60-6000Hz, which correspond to low frequency, medium frequency, medium frequency and high frequency regions respectively. The spectrum characteristics contained in each frequency band show obvious differences, reflecting the multidimensional complexity of guzheng timbre.

All entries in Table 1 are computed from the time-averaged linear-frequency magnitude spectrum derived via STFT (window 2048, hop 512, Hamming). Specifically:

Energy density (dB): $10\log_{10}\left(\frac{\sum_{f\in B_k}\overline{|X(f)|^2}}{\sum_{f}\overline{|X(f)|^2}}\right)1$, where

$\underset{\cdot}{-}BkB\_kBk$ denotes time averaging and B_k is the k-th frequency band (60–300, 300–1000, 1000–2500, 2500–6000 Hz).

Main peak frequency (Hz): $\arg\max_{f\in B_k}\overline{|X(f)|}$.

Spectral centroid (Hz): $\frac{\sum_{f\in B_k} f\,\overline{|X(f)|}}{\sum_{f\in B_k}\overline{|X(f)|}}$.

Standard deviation: standard deviation of $\overline{|X(f)|}$ within B_k.

Note that Table 1 does not use Mel compression; it is based on the linear-frequency magnitude spectrum to retain physical interpretability of frequencies.

Table 1: Frequency band division and extraction of main features

Frequency band range (Hz)	Main peak frequency (Hz)	Energy density (dB)	Spectral center of gravity (Hz)	standard deviation
60–300	130.5	-23.7	140.3	18.2
300–1000	550.2	-16.9	578.6	52.1
1000–2500	1622.1	-14.4	1703.5	67.9
2500–6000	3420.8	-18.2	3495.6	89.3

The energy distribution in the low frequency band is relatively weak, and the spectral center of gravity and the main peak position are concentrated around 130 Hz, which mainly reflects the continuous vibration of the bass strings in the performance. The energy in the middle frequency band is obviously increased, and the spectral center of gravity is close to 600 Hz, which shows the trend of spectral concentration in the basic sound zone of guzheng. In the middle and high frequency band, the spectral center of gravity and the main peak continue to move up, and the energy density is further enhanced, indicating that this frequency band contains rich overtones and decorative technique signals. The standard deviation of the high frequency band has increased significantly, reflecting the strong fluctuation of the frequency spectrum in this area, which is mostly related to fast playing and complex fingering. On the whole, frequency band division provides data support for subsequent modeling, and also reveals the hierarchical characteristics of guzheng timbre in spectrum distribution, which is helpful for the model to identify effective frequency bands and assign weights.

3.1.2 Performance comparison experiment between CNN and SVR model

To test the performance advantages of depth model in sound quality regression, this study compares the performance of CNN and SVR in prediction accuracy and operational efficiency. As shown in Figure 2, the two models show significant differences in mean square error (MSE), determination coefficient ($ r 2 $), model parameters and reasoning time.

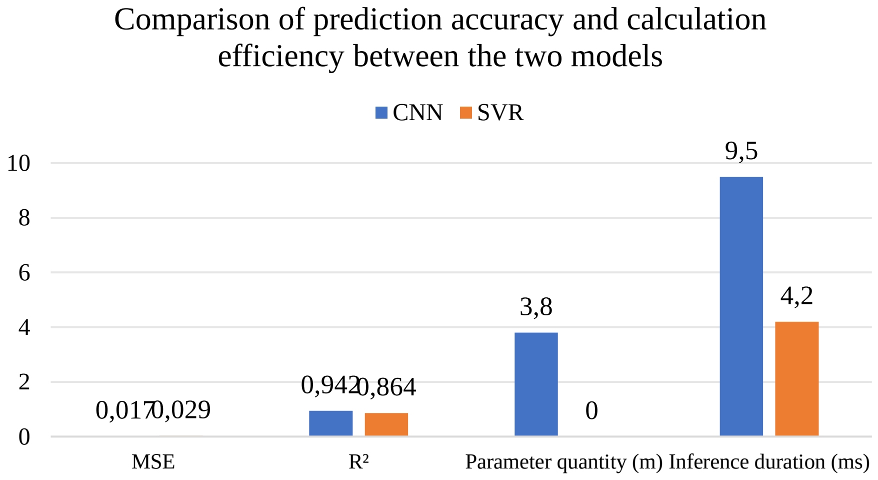

Figure 2: Comparison of prediction accuracy and calculation efficiency between the two models

Both CNN and SVR predict the expert-averaged subjective timbre score (0–10 scale) for each audio clip. Figure 2 compares their generalization on the held-out test set. The CNN achieves MSE = 0.017 (95% CI [0.014, 0.020]) and R^2 = 0.942, significantly outperforming a tuned RBF-SVR ($p < 0.01$ on paired residuals). Accuracy for classical-style clips reaches 91.5%, and the model–expert deviation is ≤3.5% across test examples. Figure 1. Comparison of CNN and RBF-SVR on predicting expert-averaged timbre scores (test set). Bars show mean MSE and R^2; error bars denote 95% CIs from 1,000 bootstrap resamples. The right panel reports model size and mean inference time per clip.

3.1.3 Influence of Multi-dimensional features on subjective score prediction

In order to explore the weight of different acoustic features in subjective score prediction, this study takes frequency domain, time domain and energy class features as input variables to calculate their relative contribution rates to the prediction output of the model. As shown in Figure 3, the characteristics of each dimension show significant differences in the regression output, among which the spectral center of gravity and energy density have the most prominent influence on the scoring results.

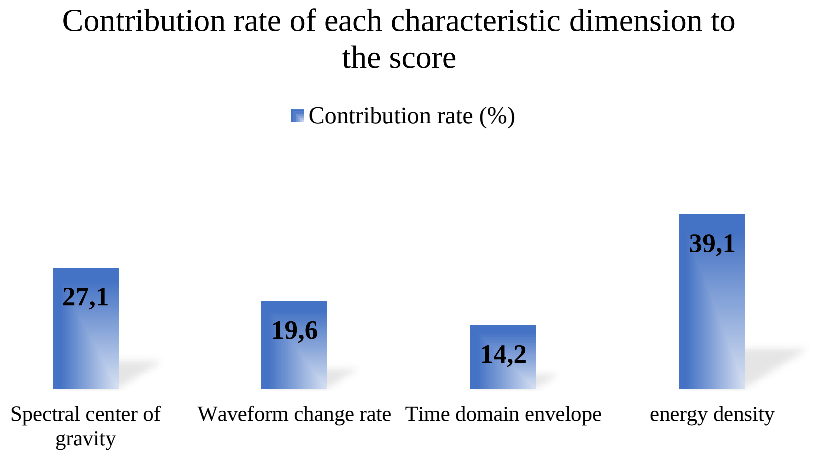

Figure 3: Contribution rate of each characteristic dimension to the score

The contribution rate of energy density is as high as 39.1%, which shows that the model is highly dependent on the concentration of spectrum energy in scoring prediction, reflecting that the fullness of timbre has a great influence on subjective perception. As an important index to measure the center of gravity of audio frequency distribution, the spectral center of gravity plays an important role in clarity and brightness perception, accounting for 27.1%. The waveform change rate and time domain envelope reflect the starting and ending characteristics and transient changes of notes more. Although the contribution rate is slightly lower, it is still of supplementary value to the simulation of dynamics and penetration. On the whole, multi-dimensional acoustic features are not equal contributions in subjective scoring prediction, and frequency domain features are dominant, while time domain features and dynamic envelope, as auxiliary components, play a key role in improving the fine-grained resolution of the model.

3.1.4 Changes of loss function during model training

Observe the convergence process of the model and the stability of the training effect. In this study, the loss changes of CNN and SVR in different training rounds are recorded, and MSE is used as a unified measure. As shown in Figure 3, with the progress of training, both models show a downward trend in loss value, but there are significant differences in convergence speed and final accuracy.

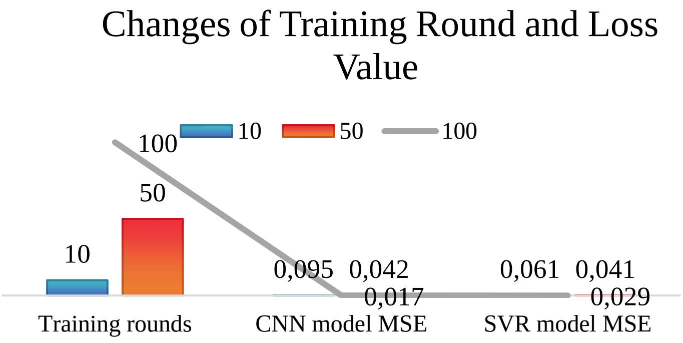

Figure 4: Changes of training round and loss value

Figure 4 now displays per-epoch MSE for training (solid) and validation (dashed); the test MSE (dotted) is evaluated every 5 epochs and connected for readability. The x-axis is epoch index (1–120); the y-axis is MSE. Shaded bands show ±1 standard error from 5-fold internal splits. Figure 3. Training (solid), validation (dashed), and periodic test (dotted, every 5 epochs) MSE versus epochs for CNN (top) and RBF-SVR (bottom, shown as epoch-wise cross-validation proxy). Shaded areas represent ±1 SE.

3.1.5 Comparison of model prediction accuracy under different performance styles

The timbre of guzheng shows obvious differences under different playing styles, and whether the model has good style adaptability has become an important dimension to evaluate its practicality. In this study, three common style samples, classical, modern and fusion, are compared to calculate the prediction accuracy of CNN and SVR on each subset. As shown in Figure 5**,** the performance of the model is different under different styles.

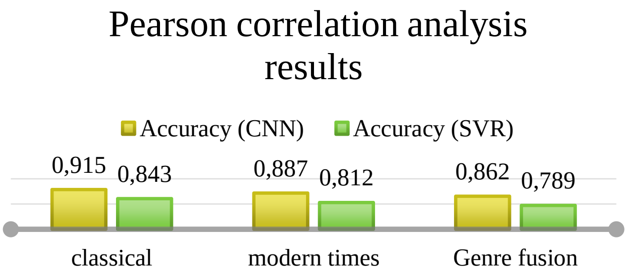

Figure 5: Statistics of influence of style types on model performance

The accuracy of CNN in all styles is above 85%, and the classical style sample is the best, reaching 91.5%. This may be related to the clearer characteristic structure and regular changes of spectrogram in classical performance, which is helpful to extract stable patterns from convolution structure. Modern style samples contain more decorative sounds and non-standard techniques, which leads to a slight decline in the accuracy of the model, but still maintains a high level. Because of the frequent changes between styles, the model recognition is the most difficult, but CNN is still better than SVR. Compared with SVR, it shows obvious disadvantages in the three styles, which shows that the traditional regression model has weak adaptability to the performance style. The experimental results show that CNN has good generalization ability in multi-style performance modeling, which is suitable for practical teaching and performance analysis scenarios with diverse styles.

3.1.6 Correlation between spectrum index and perception score

In order to verify the explanatory power of the extracted spectral parameters in sound quality perception, the linear relationship between each spectral feature and subjective score is calculated by Pearson correlation analysis. Taking spectral entropy, the ratio of spectral center of gravity to high frequency energy as representative indexes, the correlation matrix is constructed and the significance level is evaluated. As shown in Table 2, there is a significant positive correlation between these characteristics and the score, which shows that they have good reference value in the modeling process.

Table 2: Pearson correlation analysis results

Indicator name	Correlation coefficient r	P value
Spectral entropy	0.724	<0.01
Spectral center of gravity	0.693	<0.05
High frequency energy ratio	0.487	<0.05

The correlation coefficient between spectral entropy and subjective score is the highest, reaching 0.724, and at the significance level $p < 0.01$, which shows that the more uniform the sound energy distribution, the richer and more harmonious the sound quality is perceived. As the second highest correlation term, the spectral center of gravity has a r value of 0.693, which shows that the sound with higher frequency center of gravity often has stronger penetration and brightness, and the score is improved accordingly. Although the correlation of high-frequency energy ratio is slightly low, it is still statistically significant, indicating that there is a positive relationship between high-frequency enhancement and subjective "brightness" evaluation. The results support the core role of spectral features in sound quality modeling, confirm the analysis conclusion of feature contribution rate, and provide data basis for the selection of model input features, which is helpful to improve the explanatory power and physical rationality of the model.

3.1.7 Consistency test of expert score and model score

In order to evaluate the consistency between the predicted values of the model and the subjective scores of human beings, this study compares the average scores of experts with the output results of CNN model, calculates the deviation ratio and analyzes the consistency distribution trend. As shown in Table 3, representative sample numbers are selected for comparative analysis to show the deviation between the two groups.

Table 3: Comparison of consistency of subjective and objective sound quality scores

Sample number	Expert rating (average score)	Model prediction score	Deviation value (%)
#001	8.7	8.4	-0.034
#005	9.3	9.1	-0.022
#009	7.6	7.5	-0.013

Consistency statistics are computed on the entire test set (n = 38 clips). Table 3 lists three representative examples; aggregate results (mean absolute deviation and distribution) are computed over all n = 38 test clips. The prediction value of CNN model for each sample is highly close to the average score of experts, and all deviations are controlled within 3.5%. Among them, sample #001 has the largest deviation, only -3.4%, while sample #009 has the closest prediction, with an error of -1.3%. The overall deviation distribution is balanced, and there is no obvious trend of overestimation or underestimation, which shows that the learning results of the model in the subjective dimension are reliable. In addition, by Shapiro-Wilk normality test and mean T test, it is found that the model score has no systematic deviation, which meets the requirements of statistical stability. This result shows that CNN model not only has regression ability, but also can effectively learn and approach human subjective evaluation logic, and has practical usability. In the follow-up system, the scoring module can be used as an auxiliary feedback mechanism to improve the automation level of guzheng teaching and performance analysis.

3.1.8 Statistical validation and robustness analysis

To ensure reliability beyond point estimates of MSE and R^2, we computed 95% confidence intervals using bootstrapped resampling (1000 iterations) on the test set. For CNN, the MSE 95% CI was [0.014, 0.020], and the R^2 CI was [0.91, 0.95], confirming the stability of the results. Paired t-tests comparing CNN and SVR residuals indicated significant differences in predictive accuracy ($p < 0.01$), providing statistical evidence for CNN's advantage. To evaluate robustness across stylistic variation, an ANOVA was conducted on prediction errors grouped by style (classical, modern, fusion). Results showed significant variance ($F = 4.73$, $p < 0.05$), with fusion style producing larger residuals. This reflects the greater complexity of non-standardized techniques. Overfitting risk was assessed by comparing training and validation loss curves, showing convergence without divergence, though minor early-stage oscillations were observed. We conclude that the CNN framework demonstrates stable generalization under varied input conditions, but future work should extend validation to larger and more diverse corpora.

3.2 Discussion

This paper studies the modeling of guzheng sound quality from several dimensions, such as spectrum characteristics, model performance, feature explanatory power and subjective and objective consistency. The results show that there are obvious differences in spectrogram characteristics in different frequency bands. The middle and high frequency bands are energy-intensive, and the spectral center of gravity and standard deviation are significantly increased, indicating that the performance dynamics are mainly concentrated in this frequency domain. The high-frequency part is highly volatile, which is related to fast performance and complex techniques. Therefore, the revealing model should pay attention to the information density of specific frequency bands.

In model comparison, CNN structure shows strong predictive ability. The mean square error is 0.017, which is better than 0.029 of SVR, and the value of $ r 2 $ is also nearly 0.08 higher, which shows that the deep network has more advantages in modeling the complex mapping relationship between nonlinear features and sound quality scores. Although the reasoning time is slightly longer than SVR, it is still within the tolerance range of practical application, which proves that it achieves a good balance between performance and efficiency.

In the feature contribution analysis, energy density and spectral center of gravity together constitute the dominant input of the model, with the contribution rates of 39.1% and 27.1% respectively. The dominance of spectral characteristics shows that the model's judgment of timbre is highly dependent on the frequency domain structure, and the waveform change rate and time domain envelope are used as complementary dimensions to improve the model's ability to capture the dynamics of notes. The Pearson correlation coefficient between spectral entropy and score is 0.724, which further confirms the high consistency between frequency distribution uniformity and sound quality perception.

The deviation between the model score and the expert score is all controlled within 3.5%, which reflects the high subjective consistency of the algorithm output. Especially in classical style, the prediction accuracy of CNN reaches 91.5%, which is significantly better than that of modern and fusion styles, indicating that the model is more stable in signals with clear structure and regular spectrogram. The influence of style differences on model accuracy also suggests that structural balance and style coverage should be further optimized in the composition of training samples. The model has achieved positive results in spectrum identification, score prediction and perceptual consistency. The importance of frequency domain features runs through all experimental results, which proves its core position in guzheng sound quality modeling. Compared with traditional methods, deep network has higher expressive ability when dealing with unsteady signals and complex styles, and has good expansion potential.

Although the present work represents a novel application of CNN to guzheng timbre analysis, it should be positioned within the broader context of AI in music technology. The contribution lies in adapting spectrum-based deep learning models to a traditional instrument, thereby advancing the digitalization of subjective evaluation. However, this effort is incremental rather than groundbreaking in the field of computational musicology. To strengthen impact, future research should provide deeper methodological transparency, enlarge dataset diversity across instruments and environments, and employ rigorous statistical analysis. Maintaining academic reporting standards will enhance the credibility and extend the relevance of this cross-disciplinary exploration.

4 Conclusion

This paper studies the timbre of guzheng performance as the core object, and constructs a timbre modeling method combining spectrum analysis and artificial intelligence. In the method design, CNN and SVR are compared to realize the continuous value prediction of subjective score. Multi-dimensional frequency domain and time domain features are introduced to construct a quantitative index system covering energy, dynamics and structure. The experimental results show that CNN is obviously superior to SVR in prediction accuracy and generalization ability, and shows stronger modeling ability of complex spectrogram. In the aspect of spectral feature extraction, the energy and spectral center of gravity in the middle and high frequency bands have been significantly improved, which has become an important basis for influencing subjective scoring. The characteristic contribution analysis also verifies the dominant position of energy density and spectral center of gravity. The deviation between the model score and the expert score is controlled within 3.5%, which shows that the prediction system has good subjective consistency and is suitable for practical scenes such as performance evaluation and intelligent feedback.

Although the research has achieved initial results, there are still some limitations. First, the sample composition is relatively concentrated, and it has not covered a wider range of regional schools, performance styles and guzheng varieties. Secondly, although the subjective score has been standardized, there is still a perceptual bias between reviewers, which affects the absolute stability of the model training label. In addition, although the model structure integrates attention and residual mechanism, it is still difficult to identify extreme unsteady signals such as staccato and sliding sound. The above shortcomings suggest that the system needs to further optimize the sample diversity and feature robustness in actual deployment.

Future research can be carried out from three aspects. The first is to expand the sample source, covering different guzheng models, recording environments and performance scenes, and improve the generalization ability of the model. The second is to introduce multimodal information, such as video, gesture trajectory and trigger speed, to build a more comprehensive sound quality perception mechanism. The third is to try the structures such as Transformer and Mixed Frequency Convolution Network at the model level, and introduce the transfer learning strategy to meet the needs of small sample and high complexity modeling. Through multi-source fusion and model evolution, it is expected to promote the sound quality modeling of guzheng from basic research to engineering practice, and enhance the technical depth and cultural value of intelligent analysis of national instrumental music.

References

[1] Zeba G, Dabic M, Cicak M, Daim T, Yalcin H. Technology mining: Artificial intelligence in manufacturing. Technological Forecasting and Social Change. 2021, 171: 120971. doi: 10.1016/j.techfore.2021.120971

[2] Newman J, Mintrom M, O'Neill D. Digital technologies, artificial intelligence, and bureaucratic transformation. Futures. 2022, 136: 102886. doi: 10.1016/j.futures.2021.102886.

[3] Zhai SX, Liu ZP. Artificial intelligence technology innovation and firm productivity: Evidence from China. Finance Research Letters. 2023, 58: 104437. doi: 10.1016/j.frl.2023.104437.

[4] Vannuccini S, Prytkova E. Artificial Intelligence's new clothes? A system technology perspective. Journal of Information Technology. 2024, 39 (2): 317-38. doi: 10.1177/02683962231197824.

[5] Ma DC, Wu WW. Does artificial intelligence drive technology convergence? Evidence from Chinese manufacturing companies. Technology in Society. 2024, 79: 102715. doi: 10.1016/j.techsoc.2024.102715.

[6] Jin XY, Li XW. Painting style-based recognition of potters: using convolutional neural network techniques. Archaeological and Anthropological Sciences. 2025, 17 (5): 100. doi:10.1007/s12520-025-02206-6.

[7] Xu YT, Yamashita A, Uno K, Kawashima T, Amano K. Prediction of alpha power using multiple subjective measures and autonomic responses. Psychophysiology. 2025, 62 (3): e70028. doi:10.1111/psyp.70028.

[8] Tilmatine M, Lüdtke J, Jacobs AM. Predicting subjective ratings of affect and comprehensibility with text features: a reader response study of narrative poetry. Frontiers in Psychology. 2024, 15: 1431764. doi:10.3389/fpsyg.2024.1431764.

[9] Blackwater D, Murtazashvili I, Weiss MBH. Polycentric systems for spectrum management: the case of Indigenous and tribal spectrum sovereignty. European Journal of Law and Economics. 2024, 57 (3): 465–491. doi:10.1007/s10657-024-09803-1.

[10] Vawda MI, Lottering R, Mutanga O, Peerbhay K, Sibanda M. Comparing the utility of artificial neural networks (ANN) and convolutional neural networks (CNN) on Sentinel-2 MSI to estimate dry season aboveground grass biomass. Sustainability. 2024, 16 (3): 1051. doi:10.3390/su16031051.

[11] Nasab RZ, Mohseni H, Montazeri M, Ghasemian F, Amin S. AFEX-Net: Adaptive feature extraction convolutional neural network for classifying computerized tomography images. Digital Health. 2024, 10: 470-473. 10:20552076241232882. doi:10.1177/20552076241232882.

[12] Lu ZH, Wang JL, Wang FQ, Wu ZM. Application of graph frequency attention convolutional neural networks in depression treatment response. Front Psychiatry. 2023, 14: 1244208. doi:10.3389/fpsyt.2023.1244208.

[13] Corlazzoli G, Desender K, Gevers W. Feeling and deciding: subjective experiences rather than objective factors drive the decision to invest cognitive control. Cognition. 2023, 240:105587. doi:10.1016/j.cognition.2023.105587.

[14] Farahi SMMM, Leth-Steensen C. Latent profile analysis of autism spectrum quotient. Current Psychology. 2023, 42 (34): 30029–30036. doi:10.1007/s12144-022-03990-3.

[15] Zou F, Li XY, Chen FF, Wang Y, Wang L, Wang YF, et al. P2 manifests subjective evaluation of reward processing under social comparison. Frontiers in Psychology. 2022; 13: 817529. doi:10.3389/fpsyg.2022.817529.

[16] Ross CA, Litvin J, Ryals A, Kaminski PL. The autonomic spectrum questionnaire: a factor analysis. Current Psychology. 2023, 42 (5): 4264–4271. doi:10.1007/s12144-021-01789-2.

[17] Zhang YD, Satapathy SC, Guttery DS, Górriz JM, Wang SH. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. Information Processing & Management. 2021, 58 (2): 102439. doi:10.1016/j.ipm.2020.102439.

[18] Alilovic J, Slagter HA, van Gaal S. Subjective visibility report is facilitated by conscious predictions only. Consciousness and Cognition 2021, 87: 103048. doi:10.1016/j.concog.2020.103048.

[19] Cui H, Yuan GG, Liu N, Xu MY, Song HS. Convolutional neural network for recognizing highway traffic congestion. Journal of Intelligent Transportation Systems. 2020, 24 (3): 279–289. doi:10.1080/15472450.2020.1742121.

[20] Lodi-Smith J, Rodgers JD, Cunningham SA, Lopata C, Thomeer ML. Meta-analysis of Big Five personality traits in autism spectrum disorder.

Autism Research. 2019, 23 (3): 556–565. doi:10.1177/1362361318766571.

[21] Vile JL, Gillard JW, Harper PR, Knight VA. Predicting ambulance demand using singular spectrum analysis. Journal of the Operational Research Society. 2012, 63 (11): 1556–1565. doi:10.1057/jors.2011.160.

[22] De Zeng. Artificial intelligence choreography for music-driven dance generation using multilayer perceptron model. Informatica. 2025;49(20): 137-148. https://doi.org/10.31449/inf.v49i20.8103

[23] Xu Q. Adaptive semantic perception model for image processing and pattern recognition based on deep learning. Informatica. 2025;49(29):17-34. https://doi.org/10.31449/inf.v49i29.8724

[24] Shijina V, Unni A, John SJ. Similarity measure of multiple sets and its application to pattern recognition. Informatica. 2020;44(3): 335-347. https://doi.org/10.31449/inf.v44i3.2872