Real-Time Assessment of Piano Sight-Reading Ability in College Students Using CRNN

Huaning Sun
School of Music, Linyi University, Linyi 276005, Shandong, China
Corresponding author: Huaning Sun (e-mail: huaning666@163.com)

ABSTRACT This study proposes a real-time piano sight-reading assessment framework based on a multi-task convolutional recurrent neural network (MT-CRNN) for objective and intelligent evaluation of college students' performance ability. The proposed architecture integrates a shared convolutional encoder with parallel decoding branches to simultaneously perform note transcription and musical expression parameter regression. A music rule-based feature fusion module is further designed to align discrete note events with continuous acoustic features, generating a unified representation for multidimensional performance analysis. Based on this representation, an evaluation system is constructed to assess pitch accuracy, rhythmic stability, dynamic control, phrase coherence, and overall musicality. Experimental validation was conducted on a dedicated piano sight-reading dataset containing 400 performance recordings from undergraduate music students. Results demonstrate that the proposed method achieves pitch and duration F1-scores of 95.3% and 94.1 %, respectively, outperforming conventional transcription-based approaches. The generated evaluation scores exhibit strong agreement with expert assessments, with Spearman correlation coefficients reaching 0.84 for dynamic control, 0. 82 for phrase coherence, and 0.79 for overall musicality. Furthermore, the system processes an average 46.8-second performance in only 2.75 seconds, satisfying real-time feedback requirements. The proposed framework provides an effective engineering solution for automated music performance assessment by combining deep learning, acoustic signal analysis, and knowledge-guided evaluation mechanisms, offering practical support for intelligent music education systems.

INDEX TERMS piano sight-reading assessment; convolutional recurrent neural network; multi-task learning; music performance evaluation

I. INTRODUCTION

THE cross-integration of artificial intelligence and music education has brought a new paradigm to traditional skills assessment, drawing on the high-precision automated pattern recognition technologies to transform piano sight-reading evaluation from subjective teacher experience into a standardized, intelligent system [1, 2]. Developing an automated real-time evaluation system for piano sight-reading ability has both theoretical and practical significance for achieving large-scale and precise teaching, reducing teacher burden, and providing students with immediate feedback.

The core difficulties faced by the current automatic evaluation of piano sight reading are reflected in the incompleteness of evaluation dimensions and the limitations of feedback value. Most studies simplify the evaluation to a symbol comparison of audio and sheet music, which focuses on the correct or incorrect judgment of note pitch and time value [3,4]. This paradigm cannot capture and quantify the key dimensions of musical expression, such as changes in intensity, rhythm elasticity, pedal art, and coherence of musical phrases, in performance. The lack of these dimensions leads to a serious disconnect between the evaluation results and the music understanding and artistic expression abilities that universities' professional teaching aims to cultivate [5,6]. In addition, feedback generated by existing systems mostly consists of right/wrong labels or vague ratings, lacking fine-grained diagnostic information to locate the root cause of problems and guide subsequent exercises. This limitation reduces its practical value in real teaching scenarios [7,8]. Regarding visual performance evaluation, existing research mainly follows two technical paths. One is based on signal processing methods, which compare features such as the starting point and fundamental frequency of notes. These methods perform well on clear and isolated monophonic melodies, but have poor robustness to complex harmonies, legato, and pedal effects, and are completely unable to handle music performance parameters [9,10]. The second is based on deep learning methods, especially end-to-

end convolutional recurrent neural network (CRNN), which have achieved high accuracy in piano automatic transcription tasks. For example, models using conventional CRNN architectures can effectively handle the mapping problem from audio to symbol sequences [11,12]. However, these studies are essentially aimed at optimizing transcriptional performance, and their goals deviate from the needs of teaching evaluation. The high accuracy of transcription does not equal the comprehensiveness of evaluation. Existing models output discrete note events, losing the continuous expressive parameters in the sound signal. Although subsequent studies have attempted to add rule analysis after transcription, this two-stage pipeline approach is difficult to establish an intrinsic correlation between note events and expressive parameters, and the analysis is fragmented and superficial [13,14]. Therefore, the current methods have fundamental shortcomings in achieving structured and quantifiable evaluation of musical expression dimensions, which highlights the necessity of designing a new paradigm that can synchronously decode note events and expressive parameters, and conduct deep fusion analysis.

To directly address the above challenges and achieve the core goal of real-time assessment of piano sightreading ability, this paper designs and implements an improved CRNN model based on a multi-task learning framework. In this study, real-time assessment refers to providing diagnostically detailed feedback with very low latency after the completion of a short performance, typically within seconds. It is important to clarify that this system operates by processing the complete performance audio after it is recorded (fast offline processing), rather than analyzing an ongoing audio stream. This design choice prioritizes the accuracy and multidimensionality of the assessment, while still achieving the speed necessary for interactive learning. The core of this method lies in constructing a hierarchical parallel decoding structure: the main decoding branch generates note sequences through bidirectional Long Short-Term Memory (BiLSTM) and attention mechanism. Meanwhile, an auxiliary decoding branch that shares convolutional encoding features specifically regresses continuous acoustic parameters such as instantaneous intensity and spectral centroid for each time frame. The outputs of the two branches are not independent, but are time aligned and semantically integrated through a fusion module based on musicological priors, thereby generating a unified, multi-dimensional structured representation of music performance. Based on this representation, this article constructs an evaluation system that covers both technical accuracy and artistic expression.

II. MULTI-TASK CRNN PIANO SIGHT-READING MULTIDIMENSIONAL ANALYSIS METHOD FOR REAL-TIME EVALUATION

A. OVERALL FRAMEWORK OF PIANO SIGHT-READING EVALUATION SYSTEM

This paper proposes an end-to-end evaluation system based on multi-task CRNN (MT-CRNN) to address the core issue of the lack of quantification in the dimension of musical expression in real-time assessment of piano sight-reading ability. The data flow and logical relationship between the modules of the system are shown in Figure 1.

The system takes single mode playing audio as input, and outputs a structured multi-dimensional diagnostic evaluation report after feature extraction, model calculation and rule fusion. The entire process consists of four closely connected core modules: audio feature preprocessing and enhancement module, MT-CRNN encoding decoding module, music rule-based evaluation feature fusion module, and multi-dimensional diagnostic evaluation index generation module. The audio feature preprocessing module is responsible for converting the original waveform into a Mel spectrogram rich in time-frequency and dynamic information, which serves as the input representation for the model. The MT-CRNN encoding decoding module is the core of the entire system, which adopts a shared encoding and parallel decoding architecture to synchronously perform discrete classification of note event sequences and continuous regression of music expression parameters. The output of this module is fed into the evaluation feature fusion module based on music rules. Based on prior knowledge in musicology, this module precisely aligns and semantically associates discrete note events with continuous expression parameters on the timeline, generating a unified structured intermediate representation of note expressiveness. Finally, based on this intermediate representation, the multi-dimensional diagnostic evaluation index generation module derives visualized technical accuracy indicators and music performance indicators through a series of precisely defined mathematical calculations, and automatically synthesizes diagnostic text descriptions, thus completing the complete mapping from raw audio to understandable and operable evaluation feedback.

B. AUDIO FEATURE PREPROCESSING AND ENHANCEMENT

This system takes the original monaural audio waveform x(n) as input and extracts a feature representation that can simultaneously serve note recognition and music expression analysis. To achieve this goal, preprocessing abandons a single static spectrum and instead constructs an extended Mel spectrum that integrates dynamic information to enhance the model's perception of music performance related features [15,16]. Preprocessing starts with signal standardization. To ensure robust timbre analysis, we retain the full bandwidth by adopting a 44.1 kHz sampling rate. This preserves harmonics up to about 22 kHz, enabling accurate spectral centroid calculation for timbre brightness assessment while maintaining compatibility with standard music audio formats. And subjected to high pass filtering with a

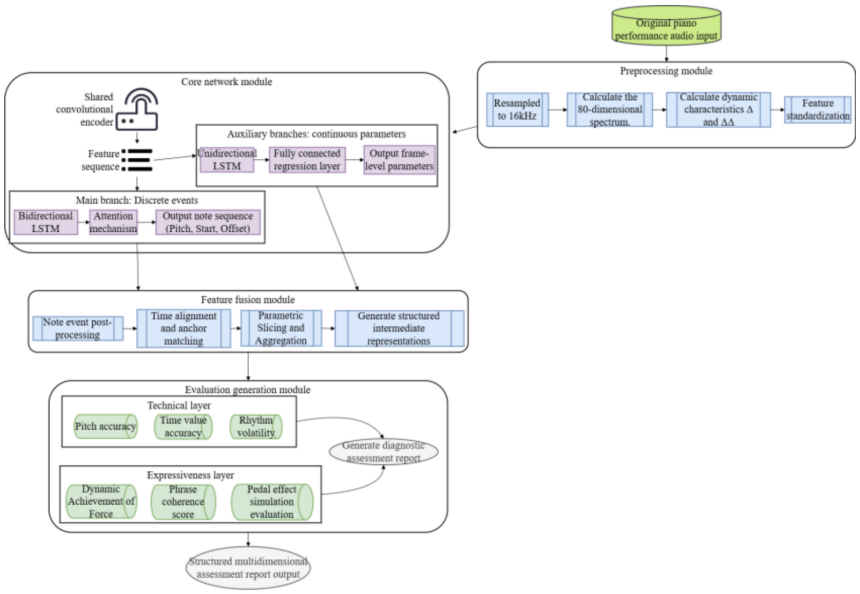

FIGURE 1. Framework of piano sight reading multi-dimensional evaluation system based on MT-CRNN

cutoff frequency of 80 Hz to eliminate lowfrequency noise. Subsequently, a pre-emphasis filter H(z)=1-αz-1(α=0.97) is applied to compensate for highfrequency attenuation and improve the clarity of harmonic structures [17,18].

The core of feature computation lies in generating and expanding the logarithmic Mel frequency spectrum. After dividing the signal into frames (with a frame length of 25 ms and a frame shift of 10 ms) and performing short-time Fourier transform, its linear spectrum is mapped to the Mel scale that conforms to human auditory characteristics through a Mel filter group consisting of 80 triangular filters, and the logarithm is taken to obtain the static feature M(t,m) [19,20]. The key here is to calculate the first-order delta and second-order delta-delta along the time axis for the static feature, forming an extended feature vector [21,22]. Before inputting into the network, the extended feature is subjected to Z-Score normalization based on the global mean μ and standard deviation σ of the training set, i.e., Mnorm=(Mext - μ)/σ, to ensure training stability. The feature representation generated by this process encodes both the static spectral contour of the pitch and the dynamic trajectory expressing the musical sense, providing a highly discriminative input basis for subsequent multi-task learning models.

C. MT-CRNN ENCODING DECODING STRUCTURE DESIGN

The core of this system design is a multi-task encoding decoding neural network structure, which synchronously maps the time-frequency feature sequences extracted from audio into discrete note event sequences and continuous music expression parameters. This structure consists of three parts: a shared convolutional encoder, a discrete event decoding branch for note transcription, and a continuous parameter decoding branch for expression parameter regression. The workflow and internal data conversion mechanism of the entire encoding decoding structure are shown in Figure 2.

The encoder receives the preprocessed extended Mel spectrogram feature Mnorm$\in$RT$\times$F (T is the time frame number, F = 240 is the feature dimension) as input. This encoder is composed of four convolution modules with the same structure stacked together, each module containing a 2D convolution layer with a kernel size of 3$\times$3, a Batch Normalization (BN) layer, and a Rectified Linear Unit (ReLU) activation function. Each convolution module is followed by a 2$\times$2 max pooling layer with a stride of 2 for spatial downsampling. Finally, the input feature map is compressed in both time and frequency dimensions to output a high-level, abstract feature sequence Henc$\in$RT'$\times$D, where T' is approximately T/16 and D is the encoding feature dimension. This encoder provides a shared underlying acoustic feature representation for the subsequent two decoding branches.

The feature sequence Henc is sent in parallel to two decoding branches with different structures. The main decoding branch is responsible for transcribing discrete note event sequences. This branch first uses a BiLSTM network to model Henc sequences, capturing contextual information before and after, and outputting a hidden state sequence H^{d}_{lstm} [23,24]. Subsequently, an additive attention mechanism is used to dynamically focus on different parts of the encoding sequence at each decoding step, and its context vector c_t is calculated by the following formula:

$$c_t = \sum_{i=1}^{T'} \alpha_{t,i} H_{\mathrm{enc},i} \tag{1}$$

The attention weight $\alpha_{t,i}$ in Formula 1 is determined by the decoder state and encoder state of the previous step. Finally, a fully connected classification layer maps the feature vector combined with attention information to

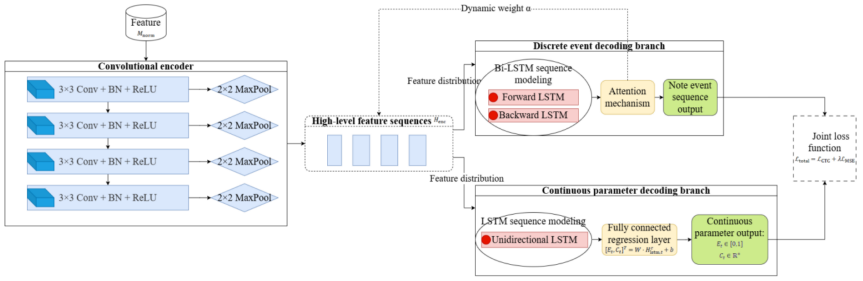

FIGURE 2. Internal working diagram of MT-CRNN encoding decoding structure

the output probability distribution at each step, covering multiple states such as no event, note start, and pitch preservation. The entire sequence is optimized through connectionist time classification loss. Parallel to the main branch, the auxiliary decoding branch is responsible for regressing continuous music expression parameters. This branch uses a unidirectional LSTM layer to process the same encoded feature sequence Henc and outputs the hidden state H^{c}_{lstm}. Its goal is to predict a continuous set of scalar values for each time frame t. This system mainly regresses two key parameters: normalized instantaneous energy $E_t \in [0,1]$, representing perceived intensity. And the spectral centroid $C_t \in R^{+}$ reflects the brightness of the timbre. The spectral centroid C_t is computed from the high-resolution spectrogram derived from the 44.1 kHz audio. This ensures that the timbre brightness feature incorporates sufficient high-frequency harmonic information, thereby providing a reliable acoustic correlate for perceptual timbre assessment in piano performance. Prediction completed through a fully connected regression layer:

$$[E_t, C_t]^T = W \cdot H^{c}_{\text{lstm},t} + b \tag{2}$$

This branch is optimized through mean square error loss. The loss functions of the two decoding branches are weighted and jointly optimized during training, and the total loss is defined as:

$$L_{\text{total}} = L_{\text{CTC}} + \lambda L_{\text{MSE}} \tag{3}$$

LCTC is a connection timing classification loss applied to the main decoding branch, specifically designed to handle sequence transcription problems where there is no strict alignment between input (audio frames) and output (note event sequences); LMSE is the mean square error loss applied to the auxiliary decoding branch to optimize the frame level regression accuracy for continuous parameters such as force and spectral centroid. λ is a hyperparameter that balances the importance of two tasks. This design forces the encoder to learn a universal feature representation that can distinguish between different notes and characterize subtle acoustic changes within the same note.

D. EVALUATION FEATURE FUSION BASED ON MUSIC RULES

The music-rule-based evaluation feature fusion aims to align and semantically associate the discrete note events generated by MT-CRNN with continuous expression parameters, generating a unified structured intermediate representation of note expressiveness. The process first performs post-processing on the original sequence of note events output by the main branch. By merging note events that overlap in time and have the same pitch, and filtering out short-term abnormal triggers with a duration below a threshold, a set of normalized note objects $N = \{n_i\}$ is generated, where each note n_i is defined by its start time t_i^s, offset time t_i^e, and pitch p_i. The core step of fusion is to map the frame level continuous parameter sequence $P(t) = [E(t), C(t)]^T$ output by the auxiliary branch to the duration interval of each note object. For each note n_i, its corresponding expression feature vector $\mathbf{f}_i$ is calculated by statistically aggregating the parameter sequences within that interval:

$$\mathbf{f}_i = \left[\frac{1}{t_i^e - t_i^s} \int_{t_i^s}^{t_i^e} E(t)dt, \sigma(E(t))_{t\in[t_i^s,t_i^e]}, \frac{1}{t_i^e - t_i^s} \int_{t_i^s}^{t_i^e} C(t)dt \right]^T \tag{4}$$

This vector sequentially contains the average intensity, standard deviation of intensity variation, and average spectral centroid of the note. Ultimately, each note object is expanded into tuples $(t_i^s, t_i^e, p_i, \mathbf{f}_i)$ to form a complete structured representation S, providing a direct and aligned data foundation for subsequent multidimensional diagnostic evaluation index calculations.

E. GENERATION OF MULTIDIMENSIONAL DIAGNOSTIC EVALUATION INDICATORS

Based on the structured note-performance sequence S generated in the previous stage, multidimensional diagnostic evaluation indicators are derived. Through a series of mathematical calculations, visualized technical accuracy and music performance indicators are derived, and finally a diagnostic evaluation report is automatically synthesized.

The calculation of technical accuracy indicators is based on the standard Musical Instrument Digital Interface (MIDI) score corresponding to the performance audio as the reference standard [25,26]. The pitch accuracy A_p and the pitch time accuracy A_d are calculated using a sequence alignment

algorithm to determine the intersection to union ratio of matching notes [27,28]:

$$A_p = \frac{|N_{\text{match}}|}{|N_{\text{perf}}|} \tag{5}$$

$$A_d = \frac{\sum_{i \in \text{match}} \min(d_i^{\text{perf}}, d_i^{\text{score}})}{\sum_{i \in \text{match}} \max(d_i^{\text{perf}}, d_i^{\text{score}})} \tag{6}$$

d_i represents the time value of a musical note. The rhythm volatility Rv is calculated based on the starting time deviation sequence $\{\Delta t_i\}$ of the matching notes to determine its standard deviation:

$$R_v = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\Delta t_i - \overline{\Delta t})^2} \tag{7}$$

The music expression index is the core of this study, aiming to quantify the level of artistic processing. The dynamic implementation degree D_f of force is obtained by calculating the Pearson correlation coefficient between the actual force curve E(t) and the ideal force contour I(t) constructed from the musical score force markers [29,30]:

$$D_f = \frac{\text{cov}(E, I)}{\sigma_E \sigma_I} \tag{8}$$

The coherence score C_p of musical phrases is calculated based on the smoothness of the expression feature vectors f_i and f_{i+1} of adjacent notes ni and n_{i+1} in the dimensions of intensity and timbre

$$C_p = \frac{1}{M}\sum_{k=1}^{M} \exp(-\|f_{ik} - f_{ik+1}\|^2/\gamma) \tag{9}$$

In Formula 9, M represents the logarithm of notes within the musical phrase, and γ represents the smoothing coefficient. The pedal effect simulation evaluation P_e analyzes the spectral centroid decay rate of bass notes during chord transitions, defined as:

$$P_e = \frac{1}{L}\sum_{l} \frac{C(t_l^s) - C(t_l^s + \delta)}{C(t_l^s)} \tag{10}$$

In formula 10, L represents the number of chord transitions, and δ represents the short-term analysis window. All indicator values, along with diagnostic descriptions located in specific sections, constitute the final evaluation report.

III. EXPERIMENTAL DESIGN AND IMPLEMENTATION

A. DATASET CONSTRUCTION AND PREPROCESSING

To verify the effectiveness of the system in real university piano sight-reading scenarios, this study constructed a dedicated dataset for university piano sight-reading evaluation. The data collection was conducted in collaboration with the piano department of a local music college, and all processes followed academic ethical standards and obtained informed consent from participants. This article recruited 40 undergraduate students majoring in piano (evenly distributed across grades) to record their first unprepared sight reading of 10 technical gradient tracks. The core benchmark annotation is independently completed by three senior experts based on a pre-established detailed scoring rubric. This rubric defines clear behavioral descriptors and scoring scales for each dimension. For example, dynamic control is assessed based on the accuracy of achieving marked dynamics, the smoothness of crescendo and diminuendo, and consistency of touch. The average pairwise Spearman rank correlation coefficient among the three experts is 0.85, ensuring high interrater reliability. This multi-expert, rubric-guided approach provides a robust and defensible gold standard for highly subjective artistic dimensions. The detailed statistical information of the dataset is shown in Table 1.

TABLE 1. Key statistical information of experimental dataset

Project	Description
Number of participants	40 undergraduate students majoring in piano (10 from each year)
Total number of audio samples	400
Average audio duration	46.8 seconds (standard deviation: 12.3 seconds)
Expert rating dimensions	Pitch accuracy, rhythmic stability, dynamic control, phrasing coherence, and overall musicality
Inter-rater reliability	The average Spearman correlation coefficient $\rho = 0.85$

The real sight-reading samples contain natural defects, requiring the model to have robustness in handling imperfect performances; The gradient distribution of participant level and track difficulty facilitates the evaluation discrimination of the testing system. Multidimensional expert ratings provide a reliable benchmark for verifying the consistency between automatic evaluation results and manual judgments.

B. MODEL IMPLEMENTATION AND TRAINING DETAILS

The model is implemented based on the PyTorch 1.11.0 framework and trained and validated on a single NVIDIA GPU. Each convolutional layer of the encoder is followed by batch normalization and ReLU activation function, and downsampling is performed using 2×2 max pooling. The discrete event decoding branch adopts a double-layer bidirectional LSTM; The continuous parameter decoding branch adopts a single-layer unidirectional LSTM, uses Adam optimizer, and applies the ReduceLROnPlateau scheduling strategy of halving the learning rate after the validation set loss plateau period. In the joint loss function, the weighted coefficient λ connecting the temporal classification loss and mean square error loss is determined to be 0.5 through grid search. The training adopts an early stopping strategy, and terminates when the validation set loss does not decrease for 15 consecutive epochs. The key hyperparameter configurations of the model are shown in Table 2.

TABLE 2. Key hyperparameter settings for the model

Hyperparameter categories	Detailed settings
Input features	80-Vimel spectrum + Δ + ΔΔ
Encoder channel	[64, 128, 256, 512]
Bi-LSTM hidden unit	512 (2 layer)
LSTM hidden unit	256 (1 layer)
Optimizer and learning rate	Adam (Initial learning rate = 0.001)
Loss weight (CTC:MSE)	1 : 0.5
Batch size	16

The hyperparameter settings in Table 2 reflect key considerations in model design. The gradual doubling of the number of encoder channels aims to build a hierarchical representation capability from local features to global semantics. The difference in the number of hidden units between BiLSTM and unidirectional LSTM reflects the different modeling requirements of the two tasks: note transcription requires stronger contextual modeling, while continuous parameter regression focuses more on the smoothness of frame level features. The setting of loss weight λ=0.5 indicates that in the joint optimization process, discrete event classification and continuous parameter regression are given approximate importance, which ensures that the encoder learns a universal feature representation that takes into account the requirements of both types of tasks.

C. COMPARISON OF BASELINE AND EVALUATION PROTOCOL

To evaluate the performance of the proposed method (denoted as MT-CRNN), this study established three baseline methods and an evaluation protocol based on expert scoring. Since multidimensional automatic assessment of piano sight-reading is still an emerging field, existing methods have significant shortcomings in the synchronous decoding and integrated assessment of pitch, duration, and expression parameters for teaching-oriented sight-reading assessment systems that require real-time multidimensional diagnostic feedback. Therefore, this paper designs a comparative scheme covering different technical paradigms to systematically verify the effectiveness of the proposed method.

1) Baseline Method 1 (OF) employs the classic Onsets and Frames framework for piano transcription. This method does not include a music expression analysis module; its transcription results are only used to calculate pitch and tempo accuracy, representing the current level of automated evaluation technology centered on note transcription.

2) Baseline Method 2 (CRNN with Rule-based Post-processing, CRNN+RP): Based on the output of the main branch (standard CRNN transcription model), this paper uses the publicly available tool librosa to extract the root mean square energy of each frame from the original audio as a velocity approximation, and performs post-association analysis (i.e., rule-based post-processing) with the identified notes through simple time alignment. This approach simulates the common practice of adding simple expression analysis to existing transcription models.

3) Baseline Method 3 (Two-Stage) employs a two-stage evaluation model. The first stage of this model uses a pre-trained transcription network, and the second stage uses an independent bidirectional LSTM network to predict the musicality score of the note feature sequences extracted in the first stage. Its design philosophy is to separate transcription from evaluation.

The core of the evaluation protocol is to measure the consistency between automatic evaluation results and expert evaluations. Specifically, on the test set, this article calculates the Spearman rank correlation coefficient between the output scores of MT-CRNN and each baseline method on five evaluation dimensions (pitch, rhythm, intensity, coherence, musicality) and the average scores of the two experts. This coefficient evaluates the consistency in ranking between two sets of data and is suitable for handling non strictly linear scoring data. The higher the correlation coefficient, the more consistent the automatic evaluation method is with the judgment of human experts in the corresponding dimension.

IV. RESULTS AND ANALYSIS

A. EVALUATION OF BASIC PERFORMANCE OF NOTE TRANSCRIPTION

As the foundation of a multidimensional visual performance evaluation system, the accuracy and robustness of note transcription directly determine the reliability of subsequent music performance analysis. The purpose of this section is to systematically verify the basic performance of the MT-CRNN model proposed in this article on the core note event recognition task, compare it with current mainstream transcription methods, and deeply analyze its performance and main error components under different playing difficulties. The technical benchmark performance of the model was evaluated by transcribing the test set audio and calculating the F1-score of key indicators such as pitch and duration. The results are shown in Figure 3.

The results in Figure 3(a) show that MT-CRNN method achieved F1-scores of 95.3%, 94.1%, and 96.0% in pitch, time value, and starting point detection dimensions, respectively, which is ahead of the baseline OF piano transcription method with scores of 93.7%, 91.8%, and 94.5%. This is mainly attributed to the introduction of multi-task learning frameworks, where shared encoders enhance the ability to distinguish acoustic features of musical notes while optimizing the expression parameter regression task. The delay dashed line corresponding to the right Y-axis indicates that the average processing time is 2.8 seconds per minute of audio, which is lower than the baseline of 3.5 seconds, reflecting the advantage of the encoder decoder integrated design in computational efficiency. Figure 3(b) reveals the specific distribution of transcription errors, with missing errors accounting for the highest proportion (38.2%). This is mainly due to the low energy and difficulty in distinguishing from background noise of the decoration sound group in fast and weak playing. Figure 3(c) shows the robustness of

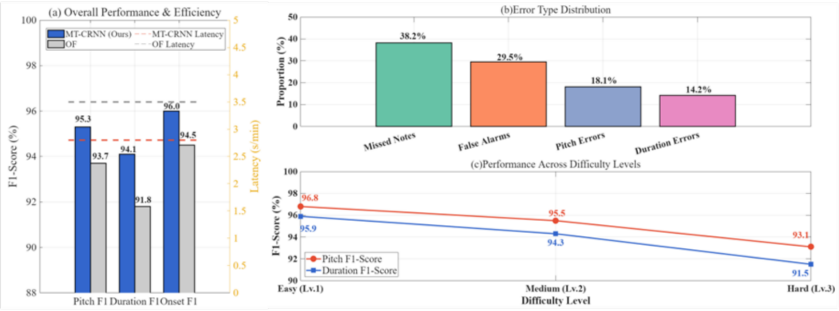

FIGURE 3. Comparison of basic performance of note transcription: (a) Overall performance and efficiency comparison; (b) Error analysis; (c) Difficulty robustness

the model on songs of different difficulty levels. The pitch and duration F1-scores show a gentle downward trend with increasing difficulty, decreasing from 96.8% and 95.9% to 93.1% and 91.5%, respectively.

B. CORRELATION ANALYSIS BETWEEN MUSIC PERFORMANCE EVALUATION AND EXPERT RATINGS

The experiment analyzed the consistency between the scores generated automatically by different methods for various dimensions and the manual grading by experienced teachers. By calculating the Spearman rank correlation coefficient between the output results of this method and three baseline methods and the expert benchmark score, the performance of each method in five dimensions of pitch accuracy, rhythm stability, intensity control, phrase coherence, and overall musicality was compared, as shown in Figure 4.

The horizontal axis in Figure 4 represents five evaluation dimensions, with the first two belonging to the category of technical accuracy and the last three belonging to the category of musical expression; The vertical axis represents the Spearman correlation coefficient ρ, with higher values indicating stronger consistency between automatic and manual evaluations. The results showed that all methods exhibited a high correlation ($\rho > 0.82$) between pitch and rhythm dimensions, which is due to the fact that these dimensions can be directly calculated from the discrete results of note transcription, and the technology is relatively mature. However, there is a significant differentiation among methods in terms of the three expressive dimensions of force control, phrase coherence, and overall musicality. The correlation of pure transcription models (Onsets & Frames) is as low as 0.25 to 0.31 due to the lack of expression analysis ability. The scheme of using rule-based post-processing to associate energy features increases the correlation to 0.45 to 0.52, which is due to rough estimation of the force parameter. The two-stage model analyzes note sequences through independent networks, and the correlation reaches 0.65 to 0.71, indicating that the separated modeling has a certain level of expressive inference ability. The MT-CRNN method in this article achieved the highest correlation in the three dimensions mentioned above, with values of 0.84, 0.82, and 0.79, respectively. This advantage is directly attributed to its end-to-end multitasking learning architecture, which forces shared encoders to learn fusion features that simultaneously encode note information and acoustic details, resulting in highly cohesive decoding of continuous expression parameters and note events in time and semantics, thus enabling more accurate simulation of experts' overall perception of music artistry. The experimental results confirm that the integrated modeling paradigm proposed in this paper is significantly superior to existing techniques such as post transcriptional processing or separate evaluation in capturing and evaluating the crucial musical expressiveness in piano sight reading.

C. VISUAL INTERPRETATION OF MULTIDIMENSIONAL DIAGNOSTIC INDICATORS

The aforementioned correlation analysis verified the effectiveness of the evaluation system from a macro statistical perspective. This section demonstrates how the system integrates multidimensional signals and musicological rules through in-depth analysis of typical performance cases, transforming abstract audio data into evaluation conclusions with clear guiding significance. The experiment selected a sight piece from the test set that includes changes in intensity and different playing techniques. The continuous parameters predicted by the system, transcribed note sequences, and derived diagnostic indicators were extracted and aligned and visualized on a unified timeline. The detailed multidimensional interpretation is shown in Figure 5.

Figure 5 uses time as the uniform horizontal axis, and Figure 5(a) uses the normalized dynamic values as the vertical axis. The two curves represent the actual dynamic curve E(t) predicted by the system and the ideal dynamic profile generated based on the score markings, respectively. The vertical dashed lines indicate the time points of key score events. At the 3.2-second mark, the actual dynamic value is 0.18, slightly lower than the ideal 0.2, which is usually caused by insufficient control of the touch force when the student starts playing a piano. The actual dynamic value increases smoothly from 0.30 at 8 seconds to 0.82 in the crescendo section from 8.5 seconds to 12.1 seconds, with relatively uniform dynamic growth control. However, at the 12-1 second mark, the actual peak value of 0.82 did not reach the ideal value of 0.9, reflecting the limitations of

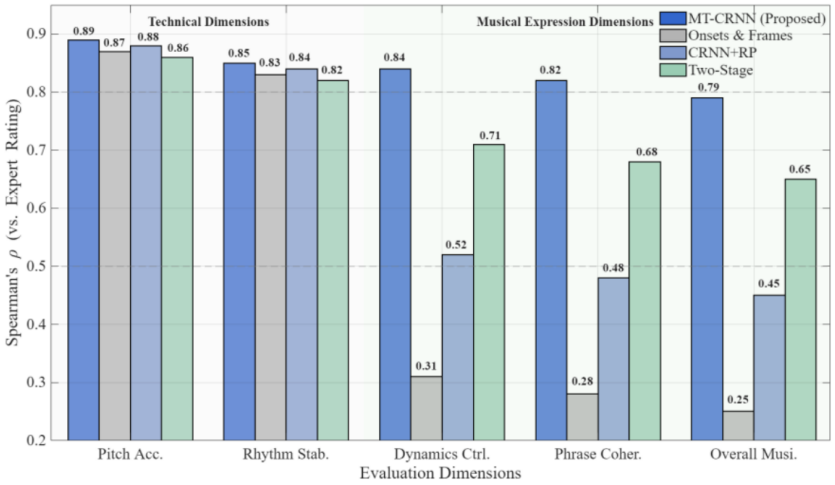

FIGURE 4. Correlation analysis between evaluation of various dimensions and expert ratings

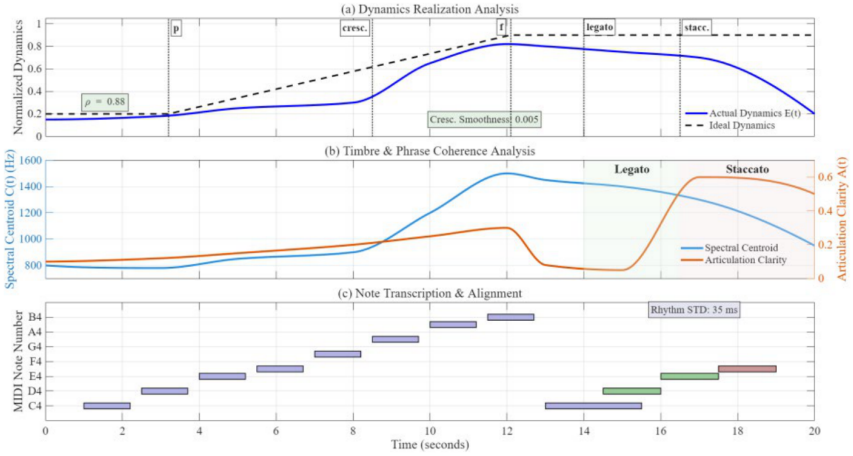

FIGURE 5. Single-sample multi-dimensional diagnostic visualization: (a) Dynamics realization analysis; (b) Timbre and coherence analysis; (c) Note transcription and alignment

students' explosive power in strong playing.

Figure 5(b) shows the analysis of timbre and pronunciation coherence, with the left vertical axis representing the spectral centroid C(t) and the right vertical axis representing the pronunciation clarity A(t). After the marked paragraph starts at 12-1 seconds, the clarity index drops sharply from 0.30 to below 0.08 and remains low. This is because the legato requires smooth transitions between notes, and the system detects a significant increase in energy envelope overlap at the beginning of the note. At the beginning of the 16.5-second staccato passage, the clarity jumps to 0.6, and the brief, separated key touch required for staccato articulation causes the acoustic features of note spacing to become abrupt.

As a hub connecting upper-level acoustic analysis with lower-level performance facts, Figure 5(c) provides an accurate spatiotemporal sequence of note events generated by the system transcription. The vertical axis represents MIDI pitch, and the horizontal position and length of the rectangular block correspond to the start and end time of each note. Starting from 14.0 seconds, the duration of the notes significantly increases and overlaps with each other (for example, the note at C4 pitch lasts for about 2.5 seconds), which physically explains the cause of the sharp drop in clarity in subgraph (b); Starting from 16 seconds, the note shape changes to short and separated (with a duration of about 1.5 seconds), directly leading to a leap in clarity indicators. This timing alignment proves that system diagnosis is not solely based on abstract acoustic parameter fluctuations, but on objective note event sequences as the cornerstone, establishing verifiable causal relationships between changes in acoustic features and specific performance techniques (such as legato and staccato). The complete analysis shows that the system can not only locate the macroscopic differences between performance and sheet music, but also quantitatively diagnose micro artistic processing dimensions such as force realization, growth smoothness, and continuous skipping execution quality by coupling timedomain parameters with note events. Its output has potential value in directly guiding targeted exercises.

D. SYSTEM STABILITY ANALYSIS UNDER DIFFERENT DIFFICULTY TRACKS

A reliable evaluation system should not only perform well under typical conditions, but also have stable generalization ability to cope with different challenges. The teaching of piano sight reading in universities covers a wide range of technical spectrum from basic to advanced, so it is crucial to verify the robustness of the evaluation system on different difficulty tracks. This section explores the correlation between performance evaluation and performance complexity. The test set samples are divided into three levels based on technical and expressive requirements: primary, intermediate, and advanced. The consistency between the system evaluation results and expert ratings is calculated for each difficulty level, and the objective acoustic and performance characteristics that cause changes in evaluation complexity are analyzed synchronously. The systematic analysis results are shown in Figure 6.

Figure 6(a) shows the Spearman correlation coefficient ρ with expert ratings on the vertical axis and difficulty level as the horizontal axis. The five trend lines indicate that the correlation of all evaluation dimensions decreases with increasing difficulty, but the decreasing patterns show differences. The technical dimensions such as pitch accuracy and rhythm stability only show a gradual decrease, from 0.91 and 0.88 in beginner difficulty to 0.86 and 0.81 in advanced difficulty, respectively. This is because note and rhythm recognition, as mature technologies, still maintains strong robustness to moderately increased note density and velocity fluctuations. In terms of musical expression, the correlation between intensity control, phrase coherence, and overall musicality has decreased more significantly, from 0.87, 0.85, and 0.83 to 0.78, 0.74, and 0.70, respectively, with an average decrease of over 10 percentage points. Figure 6(b) reveals the cause of this phenomenon through standardized analysis of multiple attribution indicators. The significantly increased density of note events and performance instability in high difficulty tracks make the estimation of continuous expression parameters face more complex signal interference; At the same time, the standardized value of expert rating dispersion increased from 0 to 1, indicating an increase in subjective differences in the artistic evaluation of high-level performances, which constitutes the objective upper limit of the automatic evaluation model approaching human judgment. Comprehensive analysis shows that this system exhibits good difficulty robustness in technical evaluation, while the performance degradation in music expression evaluation is mainly due to the increased mixing of acoustic signals in high complexity performances and the inherent subjectivity of art evaluation. This indicates an optimization direction for future models to focus on improving their analytical ability for complex music expression.

E. QUANTITATIVE EVALUATION OF EFFICIENCY IN REAL-TIME ASSESSMENT

After verifying the accuracy, diagnostic ability, and difficulty robustness of the evaluation system, its practicality ultimately depends on whether it can provide immediate feedback in actual teaching. The experiment recorded the time taken by the system to process all test set samples, calculated the total processing time, and compared it with the baseline method. At the same time, the internal process was decomposed into stages such as feature extraction and network inference for detailed time analysis, and the relationship between processing time and input audio length was analyzed. The multidimensional results of the efficiency analysis mentioned above are shown in Figure 7.

The vertical axis of Figure 7(a) represents the total processing timeTaking an audio with an average duration of 46.8 seconds as an example the results show that the average processing time of this system is 2.7 seconds, corresponding to a processing speed of 3.5 seconds per minute.The processing time of the baseline method Onsets & Frames and the two-stage model are 3.98 seconds and 4.51 seconds,respectively. Their lower efficiency is mainly due to the additional computation or serial overhead caused by the non-integrated design of the model architecture. Figure 7(b) reveals the internal time distribution of this system, with the network inference stage accounting for up to 67.3% (1.85 seconds), which is determined by the forward propagation computational complexity of the deep CRNN model; Feature fusion accounts for 16.4% (0.45 seconds), derived from rule-based alignment and metric calculation; Feature extraction and report Input/Output (I/O)account for 10.2% and 6.2 %, respectively, indicating that data preprocessing and output are not bottlenecks.

The scatter plot and fitting line in Figure 7(c) indicate a strong linear relationship between total processing time and audio length. For every 1 second increase in audio, the processing time only increases by about 56 milliseconds, and the system delay is predictable. Figure 7(d) further decomposes the trend of time consumption with audio length in each stage, where the network inference time has the largest growth slope, while the feature extraction and I/O time consumption are basically constant. This confirms that neural network computation is the core determinant of efficiency, and its growth is linear and controllable. Overall, this system has achieved better processing efficiency than real-time with its integrated architecture. Its time consumption is mainly dominated by neural network calculations and shows a stable linear relationship with audio length. This provides a solid performance guarantee for achieving second-level instant feedback in practical teaching scenarios.

V. CONCLUSION

This article proposes a solution based on MT-CRNN to address the lack of quantification in musical expression for university piano sight-reading, utilizing a framework

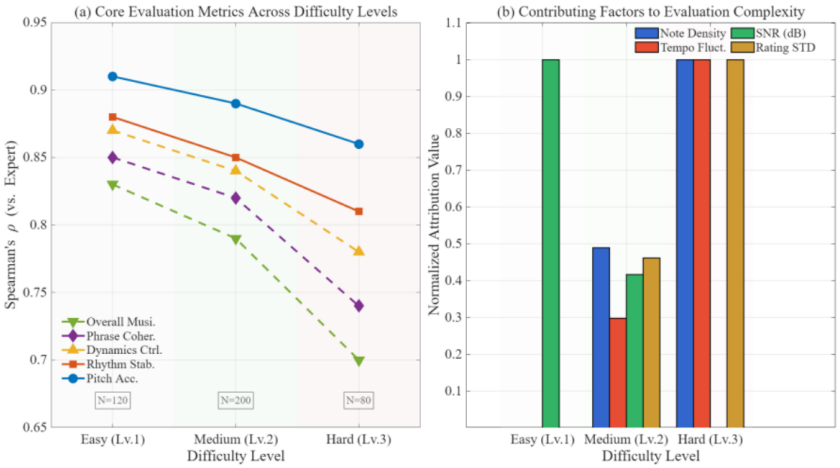

FIGURE 6. Difficulty and Robustness Analysis of System Evaluation Performance: (a) Trend of Core Evaluation Indicators; (b) Attribution Analysis of Evaluation Complexity

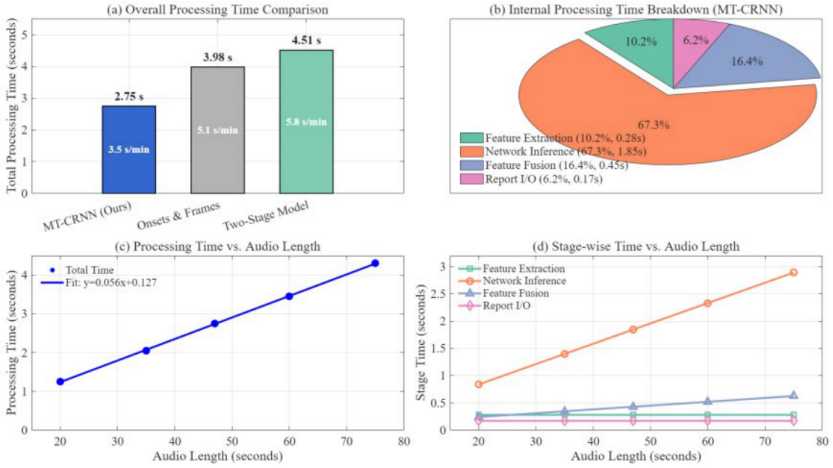

FIGURE 7. Quantitative analysis of real-time evaluation efficiency of the system: (a) Comparison of overall processing time; (b) Decomposition of internal processing time; (c) Relationship between processing time and audio length; (d) Trend of phased processing time

inspired by the automated feature extraction methods to ensure every nuanced performance parameter is objectively measured. This method is based on a MT-CRNN framework, which synchronously achieves precise transcrip-tion of note events and continuous regression of music expression parameters through a shared encoding and parallel decoding architecture. It generates a unified structured intermediate representation through a music rule-based fusion module, thus constructing a diagnostic evaluation system covering multiple dimen-sions such as pitch, rhythm, intensity, and coherence. The experimental results show that this method effec-tively achieves multidimensional and efficient evaluation goals through the improved CRNN architecture: achieving 95.3 % pitch F1-score and 94.1% hourly F1-score on note transcription basic tasks, which is superior to the Onsets & Frames baseline method. At the level of music expression evaluation, the Spearman correlation coefficients between the scores of the three dimensions of intensity control, phrase coherence, and overall musicality and expert evaluation reached 0.84, 0.82, and 0.79, respectively, significantly surpassing the baseline methods that rely on post transcriptional processing or two-stage modeling, verifying its effectiveness in quantitative evaluation in both technical and artistic dimensions. More importantly, the integrated CRNN model achieves fast processing efficiency, with the system processing an average of 46.8 seconds of visual audio in only 2.75 seconds, achieving the best processing efficiency and meeting the demand for instant feedback. This study designs and validates an end-to-end architecture based on MT-CRNN, combining deep learning technology with musicological evaluation rules. It not only achieves accurate quantification of multidimensional performance of piano sight reading, but also ensures the efficiency of the evaluation process.

AUTHOR CONTRIBUTIONS

Sun Huaning designed, collected and analyzed the data, and drafted the manuscript. Sun Huaning conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be pub-

lished. Sun Huaning participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Linyi University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] R. Rui, M. S. Amran, and N. M. Nasri, "Piano sight-reading teaching and music education: A systematic literature review," JOURNAL OF INFRASTRUCTURE, POLICY AND DEVELOPMENT, vol. 9, no. 1, Art. no. 10391, 2025, doi: 10.24294/jipd10391.

[2] N. Jia and P. Roongruang, "Training Students' Sight-Reading Skill for Piano Teaching in Colleges Under Multicultural Background in China," Journal of Modern Learning Development, vol. 8, no. 3, pp. 409-416, 2023, [Online]. Available: https://so06.tci-thaijo.org/index.php/jomld/article/view/259443.

[3] V. Phanichraksaphong and W. H. Tsai, "Automatic assessment of piano performances using timbre and pitch features," Electronics, vol. 12, no. 8, pp. 1791, 2023, doi: 10.3390/electronics12081791.

[4] X. Zhao, Y. Wang, and X. Cai, "A ResNet-based audio-visual fusion model for piano skill evaluation," Applied Sciences, vol. 13, no. 13, pp. 7431, 2023, doi: 10.3390/app13137431.

[5] J. Park, J. Kim, J. M. Park, A. Choi, W. S. Li, J. Park, et al., "Piano performance evaluation dataset with multilevel perceptual features," Scientific Reports, vol. 14, no. 1, Art. no. 23002, 2024, doi: 10.1038/s41598-024-73810-0.

[6] D. Mao and S. Liu, "Quantitative Evaluation of Piano Performance Technique and Style Based on Continuous Discretization Method," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, pp. 1, 2024, doi: 10.2478/amns.2023.2.00093.

[7] V. Phanichraksaphong and W. H. Tsai, "Automatic evaluation of piano performances for STEAM education," Applied Sciences, vol. 11, no. 24, Art. no. 11783, 2021, doi: 10.3390/app112411783.

[8] W. Wang, J. Pan, H. Yi, Z. Song, and M. Li, "Audio-based piano performance evaluation for beginners with convolutional neural network and attention mechanism," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, no. 1, pp. 1119-1133, 2021, doi: 10.1109/taslp.2021.3061267.

[9] Z. Xiao, X. Chen, and L. Zhou, "Polyphonic piano transcription based on graph convolutional network," Signal Processing, vol. 212, no. 1, Art. no. 109134, 2023, doi: 10.1016/j.sigpro.2023.109134.

[10] M. Li, "Design and implementation of piano audio automatic music transcription algorithm based on convolutional neural network," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2025, no. 1, pp. 26, 2025, doi: 10.1186/s13636-025-00412-7.

[11] J. Dai, Q. Zheng, Y. Wang, Q. Shan, J. Wan, and W. Zhang, "Multi-Feature Fusion for Automatic Piano Transcription Based on Mel Cyclic and STFT Spectrograms," Electronics, vol. 14, no. 23, pp. 4720, 2025, doi: 10.3390/electronics14234720.

[12] R. Guo and Y. Zhu, "Research on the Recognition of Piano-Playing Notes by a Music Transcription Algorithm," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 29, no. 1, pp. 152-157, 2025, doi: 10.20965/jaciii.2025.p0152.

[13] M. Zhang, "Advancing deep learning for expressive music composition and performance modeling," Scientific Reports, vol. 15, no. 1, Art. no. 28007, 2025, doi: 10.1038/s41598-025-13064-6.

[14] H. Zhang, S. Chowdhury, C. E. Cancino-Chacón, J. Liang, S. Dixon, and G. Widmer, "Dexter: Learning and controlling performance expression with diffusion models," Applied Sciences, vol. 14, no. 15, pp. 6543, 2024, doi: 10.3390/app14156543.

[15] V. Sareen and K. R. Seeja, "Speech Emotion Recognition using Mel Spectrogram and Convolutional Neural Networks (CNN)," Procedia Computer Science, vol. 258, no. 1, pp. 3693-3702, 2025, doi: 10.1016/j.procs.2025.04.624.

[16] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, "A comprehensive study based on MFCC and spectrogram for audio classification," Journal of Information and Optimization Sciences, vol. 44, no. 6, pp. 1057-1074, 2023, doi: 10.47974/JIOS-1431.

[17] J. H. Wang, P. T. Le, S. J. Kuo, T. C. Tai, K. C. Li, S. L. Chen, et al., "Audio pre-processing and beam-forming implementation on embedded systems," Electronics, vol. 13, no. 14, pp. 2784, 2024, doi: 10.3390/electronics13142784.

[18] Barkovcka OYu and A. O. Gavrashenko, "Research of the impact of noise reduction methods on the quality of audio signal recovery," Informatsiyno-Keryyuchi Cictemi na Zaliznichnomy Trancporti, vol. 29, no. 3, pp. 57-65, 2024, doi: 10.18664/ikszt.v29i3.313606.

[19] T. Kawamura, Y. Kinoshita, N. Ono, and R. Scheibler, "Acoustic scene classification using inter-and intra-subarray spatial features in distributed microphone array," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, no. 1, pp. 65, 2024, doi: 10.1186/s13636-024-00386-y.

[20] C. Asuai, A. P. Arinomor, C. T. Atumah, I. F. Kowhoro, and D. E. Ogheneochuko, "Hybrid CNN-LSTM Architectures for Deepfake Audio Detection Using Mel Frequency Cepstral Coefficients and Spectogram Analysis," American Journal of Mathematical and Computer Modelling, vol. 10, no. 3, pp. 98-109, 2025, doi: 10.11648/j.ajmcm.20251003.12.

[21] M. Swain, B. Maji, P. Kabisatpathy, and A. Routray, "A DCRNN-based ensemble classifier for speech emotion recognition in Odia language," Complex & Intelligent Systems, vol. 8, no. 5, pp. 4237-4249, 2022, doi: 10.1007/s40747-022-00713-w.

[22] S. Madanian, O. Adeleye, J. M. Templeton, T. Chen, C. Poellabauer, E. Zhang, et al., "A multi-dilated convolution network for speech emotion recognition," Scientific Reports, vol. 15, no. 1, pp. 8254, 2025, doi: 10.1038/s41598-025-92640-2.

[23] Y. Feng, "Intelligent speech recognition algorithm in multimedia visual interaction via BiLSTM and attention mechanism," Neural Computing and Applications, vol. 36, no. 5, pp. 2371-2383, 2024, doi: 10.1007/s00521-023-08959-2.

[24] Y. L. Chen, N. C. Wang, J. F. Ciou, and R. Q. Lin, "Combined bidirectional long short-term memory with mel-frequency cepstral coefficients using autoencoder for speaker recognition," Applied Sciences, vol. 13, no. 12, pp. 7008, 2023, doi: 10.3390/app13127008.

[25] F. Simonetta, F. Avanzini, and S. Ntalampiras, "A perceptual measure for evaluating the resynthesis of automatic music transcriptions," Multimedia Tools and Applications, vol. 81, no. 22, pp. 32371-32391, 2022, doi: 10.1007/s11042-022-12476-0.

[26] P. Wang and N. Dai, "Processing piano audio: Research on an automatic transcription model for sound signals," Journal of Measurements in Engineering, vol. 13, no. 1, pp. 130-139, 2025, doi: 10.21595/jme.2024.24345.

[27] K. Shibata, E. Nakamura, and K. Yoshii, "Non-local musical statistics as guides for audio-to-score piano transcription," Information Sciences, vol. 566, no. 1, pp. 262-280, 2021, doi: 10.1016/j.ins.2021.03.014.

[28] B. Bhattarai and J. Lee, "A comprehensive review on music transcription," Applied Sciences, vol. 13, no. 21, Art. no. 11882, 2023, doi: 10.3390/app132111882.

[29] G. Jones and A. Friberg, "Probing the underlying principles of dynamics in piano performances using a modelling approach," Frontiers in Psychology, vol. 14, no. 1, Art. no. 1269715, 2023, doi: 10.3389/fpsyg.2023.1269715.

[30] M. Nusseck, I. Czedik-Eysenberg, C. Spahn, and C. Reuter, "Associations between ancillary body movements and acoustic parameters of pitch, dynamics and timbre in clarinet playing," Frontiers in Psychology, vol. 13, no. 1, Art. no. 885970, 2022, doi: 10.3389/fpsyg.2022.885970.