Research on Intelligent Recognition Algorithm of Piano Playing Fingering and Generation of Personalized Practice Schemes

Jiayin Dong
School of Music and Dance, Harbin University, Harbin 150086, Heilongjiang, China
Corresponding author: Jiayin Dong (e-mail: 18445865202@163.com)

ABSTRACT Intelligent recognition of piano fingering requires fine-grained modeling of hand motion, note events, and acoustic characteristics. Existing fingering recognition algorithms often suffer from incomplete feature extraction, weak antiinterference capability, and insufficient integration with personalized practice guidance. This study proposes a multimodal fingering recognition and practice-plan generation framework based on image, MIDI, and audio data. Highdefinition video captures hand motion trajectories, MIDI data provide note position and dynamics, and audio signals supply spectral and timbral cues. HRNet extracts 21 hand key points, Word2Vec-CBOW encodes MIDI sequences, and MFCC-based features represent audio signals. A Dynamic Channel Attention module performs multimodal feature fusion , and an improved Transformer with Position-Aware Encoding captures irregular temporal dependencies in fingering sequences. A personalized practice-generation model further combines recognition outputs, learner profiles, and teaching rules to dynamically adjust practice content, difficulty, and pace. Experiments using the MAESTRO dataset and real performance data show that the method achieves 96.2% Finger Accuracy and 94.5% Direction Accuracy, outperforming HMM, Bi-LSTM, and standard Transformer baselines. After personalized practice intervention, learners' fingering standardization improves by 28.3% and practice efficiency by 32.1%. The framework supports multimodal signal processing, temporal sequence modeling, and intelligent music-training systems.

INDEX TERMS piano fingering recognition, multimodal feature fusion, transformer algorithm, audio signal processing, personalized practice

I. INTRODUCTION

PIANO performance is a complex activity involving fine motor skills and musical cognition. As the most fundamental element in piano playing, finger technique directly influences the fluency, expressiveness, and technical difficulty of performance, and it is also a critical breakthrough point for piano learners from beginners to masters. At present, the population of piano learners is expanding from adolescent beginners to adult advanced learners, who have a stronger demand for efficient and personalized practice guidance. In traditional piano teaching, finger instruction mainly relies on face-to-face demonstration and correction by teachers. Although it can provide accurate feedback, it is limited by time, geography, and teaching costs, making it unable to meet the large-scale and independent practice needs of learners. When learners practice independently, they lack objective and quantitative evaluation criteria for finger technique, making it hard to identify errors in their finger skills and leading to the formation of bad playing habits. These problems seriously affect learning efficiency and performance improvement. Therefore, researching the design of an intelligent recognition algorithm for piano playing fingering and the generation of personalized practice plans based on artificial intelligence, computer vision, and signal processing technologies can not only solve the pain points in traditional teaching but also promote the transformation of piano education from intelligence to personalization, which has important theoretical research value and practical application significance. The main innovations of this paper are summarized in three aspects. First, this paper proposes the idea of fingering recognition based on multimodal feature fusion, taking image data (hand movement trajectories), MIDI data (notes, dynamics), and audio data (key sounds) of piano performance as inputs, breaking through the limitations of unimodal data and improving the accuracy and anti-interference ability of the recognition algorithm. Second, the Transformer is enhanced by introducing Position-Aware

Encoding (PAE) and Dynamic Channel Attention (DCA) modules to optimize the ability to capture temporal features, solving the problem of insufficient recognition of the coherence of fingering sequences in piano performance. Third, a closed-loop personalized practice plan generation model of "recognition-evaluation-generation-optimization" is constructed. This paper deeply integrates fingering recognition results, learner profiles (learning level, weak points, practice preferences), and piano teaching rules to realize dynamic adjustment of practice content, difficulty, and pace, bridging the prevailing pedagogical gap between automated recognition and actionable practice guidance. The main research contents of this paper are as follows. First, a high-precision and anti-interference intelligent piano fingering recognition algorithm is designed to solve the problems of low unimodal recognition accuracy and poor temporal consistency, providing technical support for automatic fingering evaluation. Second, a personalized practice plan generation model combining learner individuality and teaching rules is constructed to integrate recognition results with practice guidance, solving the problems of homogeneity and weak pertinence of traditional practice plans. Third, experiments are conducted to verify the effectiveness of the algorithm and the scheme, providing technical solutions and experimental evidence for intelligent piano teaching, promoting the integration of artificial intelligence and music teaching, and offering a reference for the popularization and quality improvement of piano education. The rest of this paper is organized as follows. Firstly, the relevant research status is reviewed to clarify the research entry point. Secondly, the specific design of the algorithm and scheme is introduced in detail. Thirdly, the effectiveness of the algorithm is verified through experiments. Finally, the research results, shortcomings, and future research directions are summarized.

II. RELATED WORK

The design of an intelligent recognition algorithm for piano fingering and the generation of personalized practice plans require the integration of educational theory and computing technology. Sun L [1] explored the historical traces in the RCM piano score archive and revealed the evolution of teaching methods through different versions, providing a basis for version selection of training data in intelligent algorithms. Zhang Y [2] studied the current situation and reform of piano teaching in college public art education under the background of deep learning, and deep learning technology further promoted the transformation of piano fingering recognition from rule-driven to data-driven. Woosung et al. [3] proposed a personalized learning support technology based on log data of piano chord practice, and the combination of log data and fingering recognition algorithm realized quantitative tracking of practice behavior. Wang H and Zhu L [4] studied the application of XAPT-based blended learning in piano courses for non-piano majors in Qinghai area, and the blended learning environment put forward higher requirements for the cross-platform generation ability of fingering practice schemes. Yin M and Sondhiratna T [5] discussed the perspectives on piano teaching strategies at Qingdao Art School in Shandong Province, and the practical experience in teaching strategies was transformed into algorithm constraints to improve the rationality of the generated schemes. Qiu Q [6] studied a sentiment analysis method based on multimodal feature fusion, and the integration of emotional features and fingering movement features further enriched the adaptive dimension of practice schemes. Zhou K and Yan H [7] proposed MMFFNet for RGB-T crowd counting, and the multimodal fusion network structure was transferred to the fingering recognition task to enhance the joint representation ability of hand pose and keyboard space. Qin L et al. [8] proposed MFCNet for RGB-T vehicle density estimation, and the spatial distribution modeling idea in density estimation directly served the position optimization of finger landing points in fingering sequences. Yan W et al. [9] studied multi-source multimodal feature fusion in small leak detection of gas pipelines, and the anomaly location mechanism in leak detection was analogous to fingering error diagnosis to realize dynamic correction of practice schemes. Lin Y et al. [10] studied an automatic modulation classification method based on efficient multimodal feature fusion, and the efficient fusion strategy reduced the computational delay of the fingering recognition model on mobile terminals. The above studies verified the effectiveness of multimodal fusion and personalized generation in their respective fields, but lacked a dedicated fusion framework for the fine-grained movement sequence of piano fingering, and existing schemes did not integrate historical version differences and real-time log feedback into the same iterative closed-loop generation model.

III. METHOD

A. MULTIMODAL DATA COLLECTION AND PREPROCESSING

Multimodal data collection is the basis of fingering recognition. This paper collects three types of core data to ensure the comprehensiveness and accuracy of data, providing reliable support for subsequent feature extraction and model training. As shown in Figure 1, data collection equipment includes: a high-definition camera (30fps frame rate, 1920×1080 resolution) installed directly in front of the piano to fully capture the movement trajectories of left and right fingers and wrists; a piano MIDI interface to collect structured data such as note positions, key dynamics, and key durations during performance, realizing synchronous alignment with image data; a high-fidelity microphone to collect piano key audio data and capture timbre differences of keys pressed with different fingerings. Collection objects cover learners of different levels (beginner, intermediate, advanced), and performance pieces include different types such as etudes and musical pieces. Performance data of different speeds and dynamics are collected for each piece, finally constructing a multimodal dataset containing more

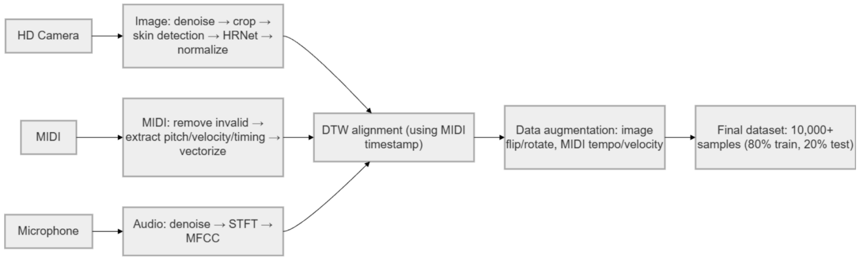

FIGURE 1. Multimodal data acquisition and preprocessing

than 10,000 performance samples, with 80% for model training and 20% for testing [11].

The core of data preprocessing is to eliminate noise, unify formats, and realize data synchronization to improve data quality. The specific processing steps are as follows. First, image data preprocessing: Gaussian filtering is used to remove image noise, hand regions are retained through image cropping, skin color detection and contour extraction techniques are used to separate hands from the background, HRNet network is used to extract 21 key points of the hand (finger joints, fingertips), and normalization is performed to map key point coordinates to the [0,1] interval to eliminate scale differences. Second, MIDI data preprocessing: redundant information is removed from the collected MIDI data, invalid notes (such as 误触keys) are eliminated, key features such as note pitch, key dynamics, key start time and end time are extracted, and converted into a standardized vector format. Third, audio data preprocessing: spectral subtraction and wavelet threshold method are used for joint denoising to remove environmental and equipment noise, short-time Fourier transform (STFT) is used to convert audio signals into spectral features, and Mel-Frequency Cepstral Coefficients (MFCC) are extracted as the core audio features. Fourth, multimodal data synchronization: To address the sampling rate disparity between the HD camera (33.3ms intervals) and MIDI events (sub-millisecond resolution), we first downsample the MIDI stream to match the video frame rate via a zero-order hold mechanism. Based on these synchronized timestamps, the Dynamic Time Warping (DTW) algorithm is employed to compensate for hardware-induced latency (typically 20-50ms), ensuring that hand-pose features, audio spectral entropy, and MIDI note-on events are aligned to a common temporal axis for joint feature fusion, ensuring that different modal data correspond to the same performance moment and avoiding time deviation affecting recognition effect. Fifth, data augmentation: image data is expanded through image rotation, flipping, brightness adjustment, etc., and MIDI dataset is expanded by changing performance speed and dynamics of MIDI data to improve the generalization ability of the model.

B. MULTIMODAL FEATURE FUSION AND FINGERING RECOGNITION MODEL CONSTRUCTION

This paper adopts a multimodal feature fusion strategy to fully mine complementary information of different modal data, and constructs a high-precision intelligent piano fingering recognition model combined with an improved Transformer architecture. First, single-modal feature extraction is performed to extract core features for three types of data respectively: hand image features are extracted by HRNet network, which can accurately capture subtle changes of finger movements through multi-scale feature fusion and output a 256-dimensional hand key point feature vector; MIDI features are extracted by Word2Vec-CBOW model, converting note sequences into vector representations, highlighting the temporal correlation between notes combined with note difference information, and outputting a 128-dimensional MIDI feature vector; audio features are extracted by STFT to obtain MFCC coefficients, and a 128-dimensional audio feature vector is constructed combined with spectral entropy, spectral centroid and other features.

Multimodal feature fusion is realized by a Dynamic Channel Attention (DCA) module, which can dynamically assign weights according to the importance of different modal features, avoid feature redundancy and improve fusion effect. The specific process is: splicing three types of single-modal feature vectors into initial fusion features $X = [F_{\text{img}}; F_{\text{midi}}; F_{\text{audio}}]$, inputting them into the DCA module. The weight coefficients W are computed via a squeeze-and-excitation mechanism:

$$W = \sigma(\mathrm{MLP}(\mathrm{GAP}(X))) \tag{1}$$

where GAP is Global Average Pooling and σ is the Sigmoid function. The final fused representation X_{fused} is obtained by $X_{\text{fused}} = W \bigotimes X$, where $\bigotimes$ denotes channel-wise multiplication, obtaining the final multimodal fusion feature vector (512 dimensions) [12].

The fingering recognition model is based on Transformer, and a Position-Aware Encoding (PAE) module is introduced to optimize the ability to capture temporal features, solving the problem of insufficient recognition of the coherence of fingering sequences in piano performance. The PAE module integrates temporal information into fusion features by embedding absolute time and relative interval features. Unlike standard Transformers that use fixed sine/cosine positional encodings $PE(pos, 2i) = \sin(pos/10000^{2i/d_{\text{model}}})$, PAE introduces a learnable temporal interval embedding Δt to capture the physical time gaps between MIDI events. The encoding for a feature at time t is formulated as:

$$E(t) = PE_{\text{absolute}}(t) \oplus \text{Linear}(\Delta t_{i,i-1}) \tag{2}$$

where $\bigoplus$ denotes element-wise addition, enabling the model to capture the precise correlation between fingerings at irregular temporal intervals; the encoder of Transformer adopts a 6-layer multi-head self-attention mechanism, each layer including a multi-head self-attention layer, layer normalization and feedforward neural network, which can fully mine deep information in fusion features; the decoder adopts a 3-layer structure, outputting recognition results of fingering types (such as finger 1, 2, 3, 4, 5) and movement directions (such as ascending, descending, horizontal). In the model training process, a joint loss function $L_{\text{total}} = L_{\text{CE}} + \lambda L_{\text{JSD}}$ is adopted. The JSD temporal consistency loss, used for enhancing the coherence of fingering sequences, is formulated as:

$$L_{\text{JSD}}(P\|Q) = \frac{1}{2}D_{\text{KL}}(P\|M) + \frac{1}{2}D_{\text{KL}}(Q\|M) \tag{3}$$

where $M = \frac{1}{2}(P+Q)$, and P, Q represent the fingering probability distributions of adjacent frames. This optimizes model parameters using the Adam optimizer with a learning rate of 1e-4 and 100 iterations, Adam optimizer is used, the learning rate is set to 1e-4, the number of iterations is 100 rounds, and early stopping strategy is adopted to avoid model overfitting.

C. DESIGN OF PERSONALIZED PRACTICE PLAN GENERATION MODEL

The personalized practice plan generation model takes fingering recognition results and learner profiles as the core (as shown in Figure 2), combines piano teaching rules to realize dynamic adaptation of practice content, difficulty and pace, and constructs a closed-loop system of "recognition-evaluation-generation-optimization". The model is mainly divided into three core modules: learner profile construction module, practice content generation module, and dynamic adaptation optimization module [13].

The core of the learner profile construction module is to quantify individual differences of learners. The specific steps are as follows: first, based on fingering recognition results, count the types and frequencies of learners' fingering errors to clarify weak points (such as non-standard cross-fingering, insufficient finger independence, uneven key dynamics); second, collect learners' practice history data, including practice pieces, practice duration, accuracy, error distribution, etc., to analyze learners' learning efficiency and practice preferences; finally, combine learners' learning goals (grading examination, hobby, professional improvement) and basic levels (beginner, intermediate, advanced) to construct a multi-dimensional learner profile, represented by vector quantization, including three core parts: weak point vector, learning ability vector, and practice preference vector.

The practice content generation module matches targeted practice content based on learner profiles and piano teaching rules. First, a practice content library is constructed containing 450 targeted exercises. This library was curated and validated by a panel of five piano pedagogy experts from the Henan Vocational Institute of Arts to ensure pedagogical soundness. Each content is marked with specific difficulty levels ($D \in [1, 10]$).

Second, the matching logic employs a Priority Scoring Function S:

$$S = \alpha \cdot \text{Weakness\_Score} + \beta \cdot \text{Ability\_Match} + \gamma \cdot \text{Preference} \tag{4}$$

The system selects exercises where S is maximized, ensuring that the most critical fingering errors (e.g., cross-fingering) are addressed first while maintaining a difficulty level within the learner's "Zone of Proximal Development" (Target Difficulty = Ability $\pm$ 1); finally, the difficulty gradient of practice content is determined combined with learners' basic levels and learning goals, ensuring that the practice content not only meets learners' current abilities but also achieves ability improvement.

The dynamic adaptation optimization module realizes real-time adjustment of practice plans to ensure gradual progress of practice. The specific process is: after learners complete practice, practice feedback data (accuracy, completion time, error distribution) are collected, combined with fingering recognition results to evaluate learners' mastery of current practice content; if the accuracy is higher than 85%, the practice difficulty is increased or switched to the next stage of practice content; if the accuracy is lower than 60%, the practice difficulty is reduced, the proportion of basic practice is increased, and the practice duration is extended; at the same time, the spaced repetition algorithm is adopted to reasonably arrange the review frequency of weak

knowledge points according to learners' forgetting curves to strengthen practice effect. Finally, a visualized personalized practice plan is generated, including practice content, practice duration, key requirements, error correction tips and progress tracking, assisting learners in efficient practice.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL ENVIRONMENT AND EXPERIMENTAL DESIGN

This experiment aims to verify the recognition accuracy, anti-interference ability of the intelligent piano fingering recognition algorithm, and the effectiveness of the personalized practice plan. The experimental environment and design are as follows. Experimental hardware environment: CPU is Intel Core i7-12700H, GPU is NVIDIA RTX 3060 (6GB), memory is 16GB, operating system is Windows 11; experimental software environment: programming language is Python 3.8, deep learning framework is PyTorch 1.12.0, data processing tools are OpenCV 4.5.5, Librosa 0.9.2, model training tool is TensorBoard.

AEM JOURNAL.ORG

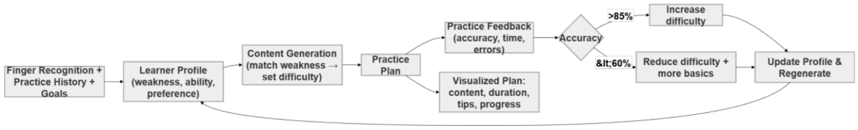

FIGURE 2. Design of personalized exercise plan generation model

The experiment is divided into two parts: the first is the performance experiment of the fingering recognition algorithm, and the second is the effectiveness experiment of the personalized practice plan. The performance experiment of the fingering recognition algorithm uses the multimodal dataset constructed in this paper, comparing the improved Transformer algorithm proposed in this paper with traditional algorithms (HMM, Bi-LSTM, pure Transformer), and the evaluation indicators include Finger Accuracy (FA), Direction Accuracy (DA), Structure Consistency Index (JSD-AS) and recognition speed; to verify the anti-interference ability of the algorithm, different interference scenarios (hand occlusion, environmental noise, performance speed change) are set to test the recognition accuracy of the algorithm in different scenarios. The effectiveness experiment of the personalized practice plan selects 60 piano learners, divided into experimental group (30 people) and control group (30 people). The experimental group adopts the personalized practice plan generated in this paper, and the control group adopts the traditional general practice plan, with an experimental period of 4 weeks. The evaluation indicators include fingering standardization, practice efficiency, and learning interest score. Among them, fingering standardization is scored by professional piano teachers according to performance videos, practice efficiency is measured by the duration and accuracy of completing the same practice content, and learning interest score is collected through questionnaire survey. This study was approved by the Ethics Committee of Harbin University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

B. ANALYSIS OF EXPERIMENTAL RESULTS

The performance experiment results of the fingering recognition algorithm are shown in Figure 3. The improved Transformer algorithm proposed in this paper outperforms traditional algorithms in all evaluation indicators: the improved Transformer algorithm achieves a recognition speed of 32fps when tested on the experimental hardware (Intel Core i7-12700H, NVIDIA RTX 3060) using a batch size of 1 and a standard input resolution of 224×224. The total end-to-end latency, including image preprocessing (approx. 8ms), feature extraction (approx. 12ms), and model inference (approx. 10ms), totals roughly 30ms. This latency falls within the threshold for real-time interactive feedback in music education; in contrast, the pure Transformer algorithm has FA of 83.0%, DA of 81.5%, JSD-AS of 80.1%; Bi-LSTM algorithm has FA of 78.3%, DA of 76.8%, JSD-AS of 75.2%; HMM algorithm has FA of 72.5%, DA of 70.3%, JSD-AS of 68.9%. The anti-interference experiment results show that in the hand occlusion scenario, the FA of the proposed algorithm drops to 89.7% and DA drops to 88.2%, still maintaining high recognition accuracy; in the environmental noise scenario, FA is 91.3% and DA is 89.8%; in the performance speed change scenario, FA is 93.5% and DA is 92.1%, all superior to traditional algorithms, verifying the anti-interference ability and generalization ability of the proposed algorithm. The effectiveness experiment results of the personalized practice plan show that there are significant differences between the experimental group and the control group after the experiment: the average fingering standardization score (on a 100-point scale) of the experimental group increases from 62.3 to 85.6 (a 28.3% improvement), while the control group increases from 61.8 to 70.2 (a 13.6% improvement). To isolate the treatment effect from natural maturation, a Difference-in-Differences (DID) analysis was performed. The results indicate a statistically significant net improvement of 14.7% attributable solely to the personalized plan ($p < 0.01$). This confirms that the proposed system provides a substantial "value-add" beyond traditional practice methods. In terms of practice efficiency, the average duration of the experimental group to complete the same practice content is 32.1% shorter than that of the control group, and the accuracy is improved by 27.8%. In terms of learning interest score, the average score of the experimental group increases from 38.7 to 48.6 before the experiment, and the control group increases from 38.5 to 41.2, indicating that the personalized practice plan proposed in this paper can effectively improve learners' learning interest. The experimental results fully verify the high precision and high reliability of the fingering recognition algorithm designed in this paper, as well as the pertinence and effectiveness of the personalized practice plan.

V. DISCUSSION

Based on the experimental results, this paper further discusses the main contents and existing problems of the above research. First, for the fingering recognition algorithm, this paper effectively solves the problems of low unimodal recognition accuracy and poor temporal consistency by using multimodal feature fusion and improved Transformer architecture. The design of PAE and DCA modules improves the model's ability to capture temporal features and the effect of multimodal feature fusion, which is consistent with the conclusions of existing relevant studies. However, this paper further improves the coherence of fingering sequences by designing a joint loss function, making the recognition results more consistent with the biomechanical constraints

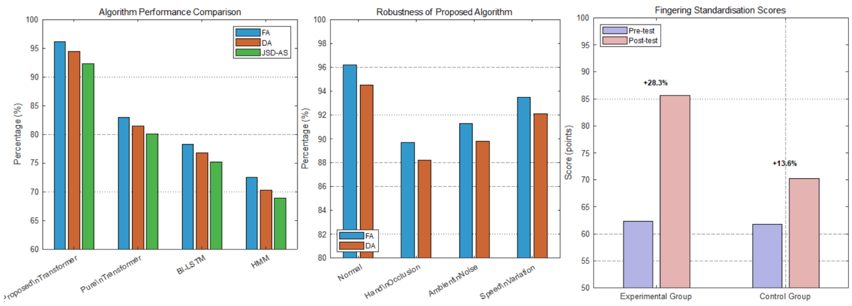

FIGURE 3. Experimental results of the performance of the finger recognition algorithm

and motor control principles of piano performance. Second, for the personalized practice plan, the joint loss function constructed in this paper can realize the deep coupling between recognition results and practice guidance, and the dynamic adaptation and optimization module can adjust practice content and difficulty according to learners' real-time practice feedback based on this coupling relationship, solving the pain points of homogeneity and weak pertinence of traditional practice plans. In the experiment, the finger technique standardization and practice efficiency of the experimental group are significantly improved, proving that this plan can help learners improve their performance level.

At the same time, the experimental results also reveal the problems existing in this research. First, the fingering recognition algorithm still has the problem of reduced recognition accuracy in severe hand occlusion scenarios (such as complete finger overlap), mainly due to insufficient accuracy of hand key point extraction. In the future, the key point extraction algorithm needs to be further optimized to improve the feature capture ability in complex scenarios. Second, the generation of personalized practice plans still relies on the preset practice content library, with certain rigidity, and the adaptability to niche pieces and personalized performance styles is insufficient, which cannot meet the special needs of some learners. Third, the experimental period is relatively short (4 weeks), and the long-term effect of the personalized practice plan has not been verified. Therefore, it is necessary to extend the experimental period in the future and further improve the dynamic adaptation logic of the plan. In addition, this research does not consider the differences in learning characteristics of learners of different age groups (such as different practice efficiency and acceptance of adolescents and adults). In the future, learner profiles can be improved to enhance the personalization of the plan.

VI. CONCLUSION

This paper addresses the problems of delayed feedback and homogeneous practice plans in finger teaching, and conducts research on the design of an intelligent recognition algorithm for piano playing finger technique and the generation of personalized practice plans through multimodal feature fusion and deep learning technologies, verifying the effectiveness of the above methods. The research results show that high-precision fingering recognition is realized by combining multimodal feature fusion with an improved Transformer structure; the closed-loop personalized practice plan based on learner profiles and fingering recognition results can improve learners' weak points, enhance finger standardization and practice efficiency, and stimulate learning interest by correcting learners' playing skills; means such as multimodal data synchronization are the key to realizing intelligent recognition and personalized guidance. The above research still has some shortcomings, such as limited recognition accuracy in severe hand occlusion scenarios, limited coverage of the practice content library, and insufficient consideration of differences among learners of different age groups. In the future, we will continue to optimize the algorithm, enrich the content library, refine learner profiles, integrate sentiment analysis, promote the development of intelligent piano education, develop an intelligent practice assistance system, and popularize the intelligent, personalized and universal development of piano education.

AUTHOR CONTRIBUTIONS

Jiayin Dong designed, collected and analyzed the data, and drafted the manuscript. Jiayin Dong conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jiayin Dong 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 was supported by,

The 2024 Harbin University Young Doctoral Research Startup Fund Project, Project Number:HUDF2024105

ACKNOWLEDGEMENTS

Not applicable.

HUMAN RESEARCH SUBJECTS

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

REFERENCES

[1] L. Sun, “Historical Traces in the RCM Piano Score Archive: Evolution of Pedagogical Approaches Through Editions,” Cultural and Religious Studies, vol. 13, no. 4, pp. 209-217, 2025, doi: 10.17265/2328-2177/2025.04.008.

[2] Y. Zhang, “Status and Reform of Piano Teaching in Public Art Education in Colleges and Universities under the Background of Deep Learning,” Archives des Sciences, vol. 74, no. s2, pp. 36-42, 2024, doi: 10.62227/as/74s25.

[3] J. Woosung, L. Eunjoo, and C. Suah, “A technique to support the personalized learning based on the log data of piano chords practicing,” The Journal of the Institute of Internet, Broadcasting and Communication, vol. 23, no. 1, pp. 191-201, 2023.

[4] H. Wang and L. Zhu, “The Application of XAPT-based Blended Learning in Piano Courses for Non-Piano Majors in Qinghai Region,” International Journal of Sociologies and Anthropologies Science Reviews, vol. 5, no. 1, pp. 395-406, 2025, doi: 10.60027/ijsasr.2025.5370.

[5] M. Yin and T. Sondhiratna, “The Perspectives On Piano Teaching Strategies At Qingdao Art School in Shandong Province,” International Journal of Sociologies and Anthropologies Science Reviews, vol. 4, no. 2, pp. 527-534, 2024, doi: 10.60027/ ijsasr.2024.4481.

[6] Q. Qiu, “Sentiment analysis based on multimodal feature fusion,” IET Conference Proceedings, vol. 2024, no. 24, pp. 497-503, 2025, doi: 10.1049/icp.2024.4566.

[7] K. Zhou and H. Yan, “MMFFNet: Multimodal Feature Fusion Network for RGB-T Crowd Counting,” IEEE Transactions on Instrumentation and Measurement, vol. 74, pp. 1-12, 2025, doi: 10.1109/TIM.2025.3605223.

[8] L. X. Qin, H. M. Sun, X. M. Duan, et al., “MFCNet: Multimodal Feature Fusion Network for RGB-T Vehicle Density Estimation,” IEEE Internet of Things Journal, vol. 12, no. 4, pp. 4207-4219, 2025, doi: 10.1109/JIOT.2024.3483175.

[9] W. Yan, W. Liu, Q. Zhang, et al., “Multisource Multimodal Feature Fusion for Small Leak Detection in Gas Pipelines,” IEEE Sensors Journal, vol. 24, no. 2, pp. 1857-1865, 2024, doi: 10.1109/JSEN.2023.3337228.

[10] Y. Lin, Q. Chen, F. Wang, et al., “Automatic Modulation Classification Based on Efficient Multimodal Feature Fusion,” Mobile Networks and Applications, vol. 30, no. 3-4, pp. 628-637, 2025, doi: 10.1007/s11036-025-02487-0.

[11] J. Hao, “Educational resource allocation optimisation driven by multimodal feature fusion,” International Journal of Information and Communication Technology, vol. 26, no. 31, pp. 105-125, 2025, doi: 10.1504/IJICT.2025.148129.

[12] E. Ka, S. Go, M. Kwak, et al., “Solar Power Generation Forecasting via Multimodal Feature Fusion (Student ABSTRACT),” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 21, pp. 23530-23532, 2024, doi: 10.1609/ aaai.v38i21.30459.

[13] H. Han, Y. Meng, X. Wu, et al., “A Transfer Learning-Based Multimodal Feature Fusion Model for Bearing Fault Diagnosis,” IEEE Transactions on Instrumentation and Measurement, vol. 74, pp. 1-13, 2025, doi: 10.1109/TIM.2025.3558745.