International Journal for Housing Science and Its Applications
Publish September 16, 2025. Volume 47, Issue 1 Pages 1042-1054

https://doi.org/10.70517/ijhsa47190

Evolution and Application of Neural Networks in High-Dimensional Data Collaborative Filtering and Piano Performance Evaluation Systems

Jia Li[1,*]
[1] School of Music, Soochow University, Suzhou, Jiangsu, 215000, China
Corresponding authors: (e-mail: Jiali_1127@sian.com).

Abstract Since the traditional deep neural network cannot provide accurate serialized recommendation in the process of high-dimensional data collaborative filtering recommendation, thus this paper adds the attention mechanism to the traditional deep neural network to optimize the model structure, and proposes the data recommendation algorithm based on the improved deep neural network. Combined with the index performance comparison between the improved algorithm and the classical serialized recommendation algorithm, the accuracy of the improved deep neural network in high-dimensional data recommendation is verified. In order to realize the multifaceted evaluation of piano playing, pitch, duration, fingering, depth, and muscle stiffness evaluation indexes are proposed. Utilizing the computational volume advantage of the deep separable convolutional algorithm and equipped with the SE Block Attention Module, the multidimensional data generated during piano playing is analyzed. The combined recognition rate of piano playing gestures (down-finger, through-finger, across-finger, expanding and retracting) by the neural network model utilizing the deep separable convolutional network architecture + SE Block attention mechanism is 93.54%, which meets the requirements of the piano playing evaluation system.

Index Terms deep neural network, depth separable, attention mechanism, high dimensional data, piano playing

I. Introduction

With the continuous development of the concept of education and teaching, students' music education gradually develops from the traditional singing teaching to the direction of diversification and synthesis. Among them, piano playing, as an important form of musical expression, can not only effectively enhance students' musical perception and performance ability, but also promote students' in-depth understanding of music theory [1], [2]. The practice and application of piano performance in music classroom teaching not only enriches the teaching means, but also stimulates students' strong interest in music and unlimited potential [3]. By listening to and imitating the teacher's piano performance, students can gradually master the basic knowledge and skills of music, and improve their music appreciation and performance ability. At the same time, the teacher in the classroom combined with the theory of piano education and for the problems arising in the performance of the student simulation teaching way to teach, so that students understand the basic tasks of piano teaching, for music education to provide an effective teaching path.

In the traditional piano performance evaluation, teachers often check the final learning outcomes of students from a piece of work played by students in a professional way, but do not have a certain learning progress evaluation [4], [5]. Therefore, based on the traditional piano course evaluation, students are prone to laziness in the student process, in order to lead to the lack of serious practice in the process of learning, and directly practicing the final exam works. It is always known that course evaluation includes not only summative evaluation but also process evaluation [6]. However, in the process of evaluation, due to the inconsistency of the evaluation standards of the teachers, the students' performance assessment in different examination rooms is also by the difference, which will also lead to the phenomenon of unbalanced performance evaluation [7], [8]. The learning ability assessment obtained by traditional evaluation methods is not objective to a certain extent, so there is an urgent need to solve the current piano performance evaluation problems through modern advanced digital technology [9], [10]. The emergence of automated assessment system for piano playing can provide students with more accurate assessment results, which is conducive to enhancing students' intuitive understanding of music [11].

In this paper, we correct the attribute features of high-dimensional data and utilize adaptive algorithms for high-dimensional data attribute feature extraction. Obtain the value of weight information between high-dimensional data information and calculate the degree of recognition between high-dimensional data. Add the attention mechanism in the deep neural network and propose the data recommendation algorithm based on the improved deep neural network. Build a comparative experimental environment to compare the data recommendation algorithm based on improved deep neural network with the classical serialized recommendation algorithm to verify its effectiveness in high-dimensional data processing. Analyze the multi-dimensional data generated by the microphone part and camera part in the piano performance evaluation process by using the depth separable convolution and SE Block attention module, and build a piano performance intelligent evaluation system. Analyze and discriminate the processing performance of depth-separable convolution and SE Block Attention Module.

II. Deep neural network-based serialized recommendation algorithm for high-dimensional data

II. A. Collaborative Filtering Recommendation Algorithm Design for High-Dimensional Data

High-dimensional data is a dataset with a large number of variables (features) and samples with dimensions far beyond what one can intuitively observe and process [12], [13].

In practice, these datasets usually contain more variables than the number of observable samples, making data processing and analysis extremely complex. In piano performance, due to different piano repertoire as well as player habits, playing styles, etc., different playing venues as well as the combination of piano playing with strings, etc. all affect the piano playing effect, the resulting data such as audio, pitch, and duration are too much and difficult to process.

Since the information volume of high-dimensional data is huge and far beyond people's intuition, a large amount of computational resources and time need to be consumed to process these data during data processing and analysis. In addition, dimensional catastrophe is a problem that needs to be faced with high-dimensional data. As the dimensionality of data increases, the complexity of data processing and analysis also increases, which may lead to performance degradation.

The challenges posed by high-dimensional data are multifaceted, and dimensional catastrophe is a very serious challenge for high-dimensional data analysis. As the dimensionality of the data increases, the complexity of data processing and analysis also increases, thus making the data analysis less efficient or error-prone. Finding and utilizing the laws of data effectively in high dimensional space is an important task in high dimensional data analysis. And many existing data analysis methods may not be applicable to high-dimensional data, and new analysis methods and techniques need to be developed, which include new data dimensionality reduction techniques, feature selection methods, visualization techniques, algorithms, and so on.

II. A. 1) High-dimensional data attribute feature extraction

Based on data mining for the extraction of high-dimensional data attribute features, firstly, the high-dimensional data are collected and transmitted through the bus, and the high-dimensional data attribute architecture diagram is constructed.

Through the data mining algorithm, the preference correction of the attribute features of the high-dimensional data is carried out and the extraction of the attribute features of the high-dimensional data is realized by using the adaptive algorithm, and the following formula is used to represent the preference value of the attribute features of the high-dimensional data:

$$C_{uu}^{*} = \sqrt{\frac{d_{in}(v_1)}{d_{out}(u_1) + d_{in}(v_1)}} \tag{1}$$

where $d_{out}(u_1)$ is the high-dimensional data attribute feature value. u_1 is the collection of associated information of the collected high-dimensional data attribute features. $d_{in}(v_1)$ is the preference value in the process of high-dimensional data feature mining, and v_1 denotes the collection of nodes of high-dimensional data attribute features.

The preference of information attribute features of the high-dimensional data attribute feature set u_1 is affected by the noise N_u. If the high-dimensional data attribute features based on the data mining rule u are represented as v, the preference features C_{uu}° of the mined high-dimensional data attribute features are pooled in, and this relationship is represented by the following equation:

$$\overline{R}_{ik} = \sum_{u_i \in v} C_{uu}^{*} R_{jk} \tag{2}$$

where, $\overline{R}_{ik}$ denotes the preference prediction of the mining result u_i for the high-dimensional data v_j, and R_{ji} denotes the preference scoring of the mining result u_i for the high-dimensional data attribute feature v_k. The prediction scoring of the attribute features of the high dimensional data is accomplished by the following equation:

$$\overline{R} = T\overline{R}_{ik} \tag{3}$$

where T denotes the weight information of the high-dimensional data attribute features.

Substituting the predicted scores obtained from Eq. (3) into the data mining algorithm, the high-dimensional data attribute feature vector is obtained. Relying on the obtained high-dimensional data feature vector, the data mining classification tree of high-dimensional data attribute features is obtained:

$$p(R \mid \sigma_R^2) = \prod_{u_j=1}^{r} \left[N(R_{ij} \mid \mu, \sigma_R^2) g(x) I_{ij}^R (U_i^T V_\lambda) \right] \tag{4}$$

where $N(R_{ij} \mid \mu, \sigma_k^2)$ denotes an attribute feature in the high-dimensional data attribute feature r obtained using data mining techniques, the standard variance of the high-dimensional data attribute feature is σ_R^2, and I_{ij}^R denotes a feature extraction function, and if the feature attribute of the high-dimensional data information U_i^T is V_λ, its preference feature is obtained.

The extraction of high-dimensional data attribute features is accomplished by mining the high-dimensional data attribute features.

II. A. 2) Similarity calculation between high-dimensional data

In the process of solving the similarity between high-dimensional data, the space vector method is mainly used to determine the value of weight information between high-dimensional data information.

The formula for calculating the weight of high-dimensional data information is as follows:

$$Weight(t)_d = \left[\alpha(t)_d + \beta(t)_d \right] \times TFIDF(t) \tag{5}$$

where $Weight(t)_d$ denotes the weight of a data information t in the high-dimensional data d, and the initial and ending weights of the data information t in the dataset d are $\alpha(t)_d$ and $\beta(t)_d$, respectively. For the TF-IDF value of the data information t in the dataset d, it can be computed by Eq. (6), viz:

$$TFIDF(t)_d = \frac{TF(t)_d \times \log 10\left(\frac{N_k}{n_i} + 0.01 \right)}{\sqrt{\sum_{xd} \left[TF(t)_d \times \log 10\left(\frac{N_k}{n_i} + 0.01 \right)^2 \right]^2}} \tag{6}$$

In the above equation, N_k denotes the amount of features in the information set of high-dimensional data, $TF(t)_d$ denotes the existence probability of high-dimensional data t in data set d, and n_i denotes the number of all high-dimensional data in data set d.

The similarity between high-dimensional data is defined as by adaptive analysis method:

$$Sim(s_i, s_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_i|} \tag{7}$$

In the above equation, C represents the set of feature information between high-dimensional data, and the similarity between high-dimensional data can be determined by Equation (7) to take values between 0 and 1.

By calculating the similarity of high-dimensional data. The distribution of similarity between different high dimensional data in the dataset is obtained, which is represented by using the following matrix:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nm} \end{bmatrix} \tag{8}$$

where s_{nm} represents the similarity between different high-dimensional data s_n and s_m in the dataset.

By analyzing the similarity between different high-dimensional data in the dataset, the high-dimensional data similarity weight threshold value is set. Through the calculation of the weight value, the similarity between the high-dimensional data in the data set is obtained, and the distribution of the similarity between the high-dimensional data information is obtained to complete the calculation of the similarity value between the high-dimensional data.

II. B.Architecture of deep neural networks

The main idea of Deep Neural Networks (DNNs) is derived from the millions of interconnected neurons in the human brain [14]. In addition to labeled training data, they have the amazing ability to automatically identify and extract relevant high-level features from raw inputs. In recent years, the increased availability of large datasets, as well as specialized hardware and efficient training algorithms, have enabled DNNs to outperform current human performance in many application areas.

A DNN model is composed of multiple layers, each containing multiple neurons. A neuron is an independent computational unit within the DNN that applies an activation function to its inputs and passes the results to other connected neurons.The more common activation functions in DNNs include sigmoid, hyperbolic tangent function, and ReLU (rectilinear unit). Typically, a DNN has at least three layers (and often more), i.e., an input layer, an output layer, and one or more hidden layers. The neurons in each layer have direct connections to the neurons in the next layer. Moreover, the number of neurons in each layer and the connections between them vary significantly from one DNN to another. Thus, overall, a DNN can be mathematically defined as a multi-input, multi-output parametric function F, consisting of multiple parametric subfunctions representing different neurons.

Each connection between neurons in a neural network has a weight parameter which represents the strength of the connection between neurons. For supervised learning, it learns the weights of the connections by minimizing the cost function of the training data.DNNs can be trained using different training algorithms, but gradient descent using backpropagation is by far the most commonly used training algorithm for DNNs.

II. C.Deep neural network based data recommendation algorithm modeling

In the process of performing high-dimensional data recommendation, the number of layers of the hidden layers of the traditional DNN network and the number of neurons corresponding to each hidden layer cannot be set precisely, which directly affects the accuracy of the entire serialized recommendation of high-level data. The traditional DNN is improved accordingly by introducing the attention mechanism for processing the piano playing dataset to pave the way for the subsequent research.

(1) Model Input Layer

In the process of serialized recommendation for high-dimensional data, the data categories are first divided into continuous features and category features. In the process of dimensionality reduction of the data, the two categories of features are mapped to the corresponding low-dimensional feature space respectively. In processing the category features, one-hot is used for encoding. When dealing with continuous features, they are directly mapped to the low-dimensional embedding space. For example, the category features in a set of samples are month, gender, and city. The one-hot is used to encode them as:

$$\begin{cases} month = 1 = [1,0,0,\cdots,0] \\ Gender = female = [1,0] \\ city = NanJing = [0,1,0,\cdots,0] \end{cases} \tag{9}$$

(2) Dense feature embedding layer

The embedding layer mainly converts the high-dimensional sparse feature vectors into low-dimensional dense features.

The expression of the corresponding mapping is:

$$H^{(0)} = [h_1, h_2, h_3, \cdots, h_m] \tag{10}$$

where $h^{(0)}$ is used as the input to the multichannel attention mechanism layer and the DNN network layer, respectively. h_i denotes the i rd vector in the feature domain, and m denotes the number of vectors in the feature domain.

(3) Multi-channel attention mechanism

The multi-channel attention mechanism layer takes the vector $H^{(0)} = [h_1, h_2, h_3, \cdots, h_m]$ obtained from the embedding layer as input, and the traditional attention mechanism weight is calculated as:

$$a = soft\max(w_{s2} tanh(W_{s1} H^T)) \tag{11}$$

where W_{s1} is the matrix of $d_a \times 2u$. w_{s2} is the dimension d_a vector. In the multi-channel attention mechanism, w_{s2} is extended to a matrix of $r \times d_a$. r is the number of channels. Multi-channel attention mechanism weights are calculated as:

$$y_a = soft\max(W_{s2} tanh(W_{s1} H^T)) \tag{12}$$

The full feature vector of the input data is defined as:

$$M = y_a H \tag{13}$$

where M is the matrix of $r \times 2u$.

(4) DNN network layer

DNN network is a kind of fully connected feedforward neural network, which consists of cross layer and depth layer, and mainly accomplishes the extraction of highly nonlinear features. The corresponding forward propagation process of the depth layer when using the $H^{(0)}$ matrix as input is:

$$y^{l+1} = \sigma(W^l H^l + \theta^{l,l+1}) \tag{14}$$

where l is the number of hidden layers and $y^l + 1$ denotes the output of layer l . W^l is the matrix corresponding to layer l , $\theta^{l,l+1}$ is the threshold between layers l and $l+1$, and the final output of the deep network can be expressed as:

$$y_{dnn} = \sigma(W^{h+1} H^h + \theta^{h,h+1}) \tag{15}$$

The transfer relationship between the layers of a cross neural network in a DNN is:

$$x_{m+1} = f(x_m, w_m, b_m) + x_m = x_0 x_m^T x_m + b_m + x_m \tag{16}$$

Assuming a b_m threshold of 0, the computational expression for the cross network between the input layer and the second layer is:

$$x_1 = x_0 x_0^{'} w_0 + x_0 \tag{17}$$

And so on, the computation of the cross-layer can be completed. Finally the output in the DNN network can be obtained by summing:

$$y_{deep} = y_{dnn} + x_{m+1} \tag{18}$$

(5) Model output

The low-dimensional features acquired through the multi-channel attention mechanism and the nonlinear features acquired through the deep network are summed to finally realize the fusion and extraction of data features. The final output of the improved deep network is:

$$y = sigmoid(y_{deep} + y_a) \tag{19}$$

II. D.Analysis of deep neural network based recommendation algorithms

II. D. 1) Experiment-related settings

The experimental configuration environment for this chapter is shown in Table 1.

Table 1: Experimental environment configuration

Environment name	Parameter information
Operating system	Windows 11
GPU	RTX 3080
CPU	Xeon® Platinum 8255C
Memory	60GB
Development tool	Visual Studio Code
Programming language	Python3.8
Depth learning framework	TensorFlow

Two publicly available datasets are used in this chapter to validate the model performance.

(1) MovieLens 1M dataset. The data contains more than 6000 users, 3416 movies, and 1 million ratings.

(2) Recommender System Competition's News Recommendation dataset. The data contains more than 1 million users and more than 60 million clicks. The dataset is real data from industry, thus giving this dataset rich and diverse exploration possibilities and experimental space.

In this chapter, the data recommendation algorithm based on improved deep neural networks is compared with the classical sequential recommendation models that have been validated and recognized in related fields, in order to test the recommendation effect and the degree of improvement of the models. The following comparison models are chosen.

FM model: the FM model combines the advantages of linear model and matrix decomposition, and has been widely used in the fields of recommender systems, advertisement click rate prediction, and so on.

DSSM model: a semantic model based on deep neural network proposed by Microsoft.

YouTubebeDNN model: learning user and item embeddings with deep neural networks.

MIND model: uses a label-aware attention mechanism based on scaling clicks to guide the learning of multi-interest expression embeddings for users.

Comirec model: uses a multiple interest module to capture multiple interests from a sequence of user behaviors and can retrieve candidate sets of items in a large-scale pool of items.

SDM model: uses gated fusion module to effectively combine long-term preferences with current shopping needs, which is more accurate than the previous direct splicing of long and short-term interests.

In this paper, HitRate@K and Recall@K are chosen as evaluation metrics, the former focuses on measuring the ability of the recommender system to capture users' immediate interests. The latter focuses on whether the system is able to cover a wide range of users' real interests. The combination of the two can effectively and succinctly evaluate the recommendation capability of the recommender system. The higher the result value, the better the recommendation effect.

II. D. 2) Results and Analysis

The results of the experiments comparing the data recommendation algorithms based on improved deep neural networks with the comparison models on the unified dataset are shown in Fig. 1. The figure shows the HitRate@50 metric scores of several algorithms on MovieLens 1M dataset with binomial symmetric distribution of graphical curves.

The hit rates of FM, DSSM, YoutubeDNN, MIND, Comirec, SDM, and the model in this paper are 22.85%, 30.17%, 29.64%, 25.32%, 30.59%, 45.63%, and 60.59%, respectively.

The click-through rate of the data recommendation algorithm based on improved deep neural network is 60.59%, which is 37.74% higher for this model compared to the traditional FM model, which may be due to the fact that the FM does not utilize deep learning, which may not provide sufficient expressive power when dealing with complex and deep user-item interaction patterns.

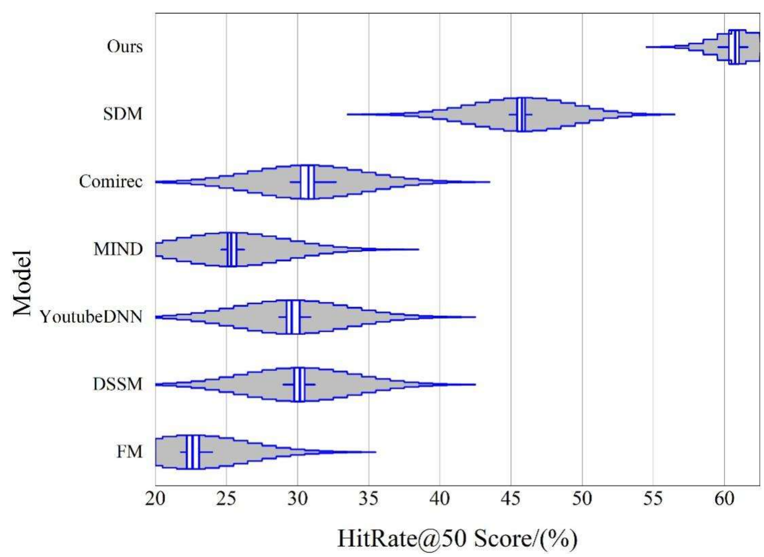

Figure 1: Comparison experiment results

The results of the comparison experiments of multiple models on the same dataset are shown in Fig. 2, in which the data curves are binomially symmetric.The click-through rate of this model is 72.08% when taking the top 150 items that users are most interested in in the MovieLens 1M dataset. Compared with the FM model and DSSM model, the click rate of this model is increased by 21.19% and 18.61% respectively. Compared with the SDM model, which takes into account the long and short-term interests, the click-through rate of this model is less improved, but still has advantages.

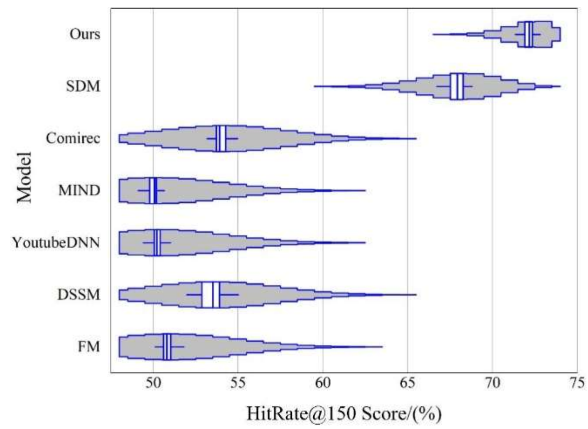

Figure 2: Comparison results of multiple models

The results of the comparison experiments of the model on the news recommendation dataset are shown in Fig. 3, and the curves have a Kernel Smooth distribution. In the news recommendation dataset, the click-through rate of the HitRate@50 indicator of this model is 0.151%.

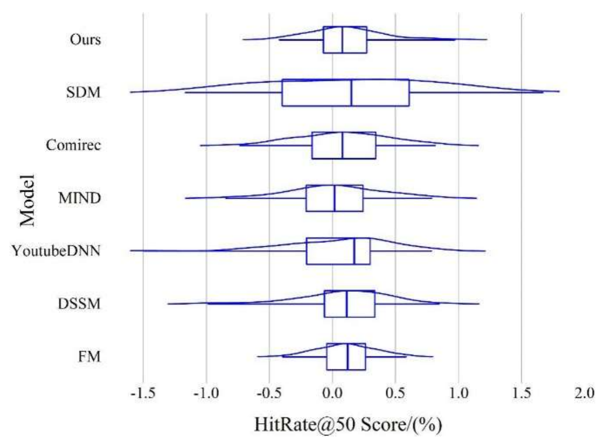

Figure 3: Comparison results of the model in the news recommendation data set

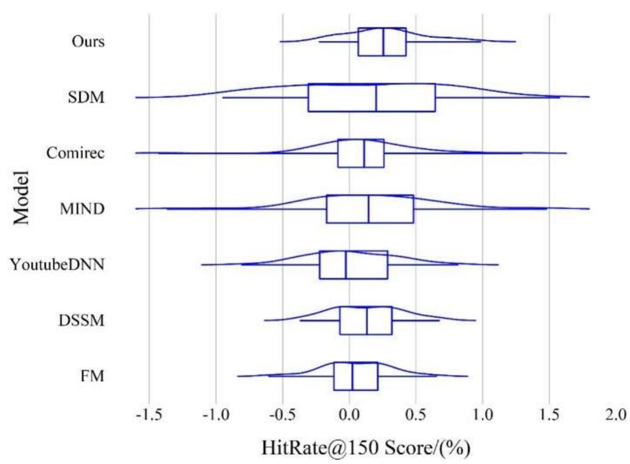

Figure 4: The comparison results of the news recommendation data concentration

When taking the top 150 items that users are most interested in, the comparison results of multiple models in the news recommendation dataset are shown in Figure 4. The click rate of this paper's model is 0.279%, which is significantly higher than other models. After analyzing the experimental data, it can be clearly seen that in this dataset, the recommendation effect of this model shows some advantages compared to most of the same type of sequence recommendation models.

Given that the actual performance values of the recall metrics of each model on the news recommendation dataset are found to be low through experiments, it is difficult to conduct an intuitive and statistically significant comparative analysis. Therefore, this paper does not provide a detailed quantitative comparison of the recall performance between the models on the news recommendation dataset for the time being.

The results of the comparison experiments of the Recall@20 metric on the MovieLens1M dataset are shown in Fig. 5, with a binomially symmetric distribution of recall curves. The score of the Recall@20 index of the proposed model is 21.43%, which is slightly lower than that of the SDM model (22.47%).

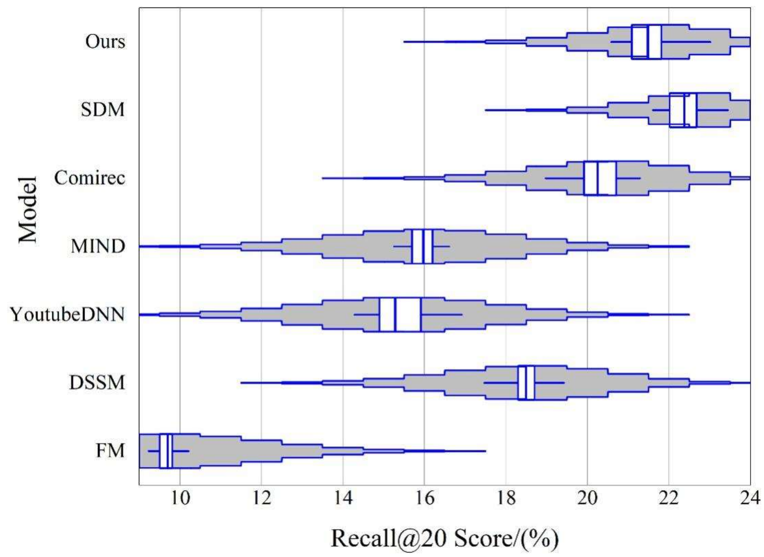

Figure 5: The comparison results of the recall index in the movielens1m data set

The results of the comparison experiments of the Recall@30 metric on the MovieLens1M dataset are shown in Figure 6. The recall curves of the algorithms have binomial symmetric distributions, and the respective recall rates are 15.69%, 27.85%, 21.01%, 21.33%, 25.62%, 32.86%, and 28.53%. The recall of the model in this paper is lower than that of the SDM model with a difference of 4.33%.

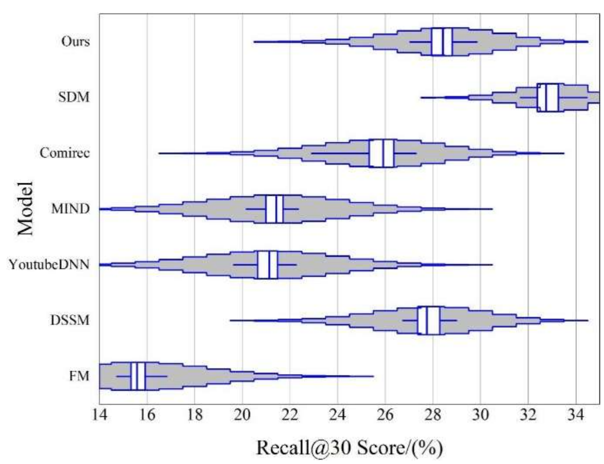

Figure 6: Compare the results in the movielens1m data set

III. Complex Semantic Scale User Context Awareness Based on High-Dimensional Data

The inclusion of multidimensional information in complex semantic contexts allows for a more comprehensive understanding of the user's behavior and intentions, thus enabling the end device to provide more accurate and personalized services.

The proposed system in this chapter aims to collect data from a smartwatch to perceive the user's piano playing process, thus assisting the piano player's learning.

III. A. Piano Performance Evaluation Indicators

In order to achieve an accurate digital mapping of the piano playing process, three professional pianists were invited to provide background knowledge and conduct preliminary research. Based on the knowledge shared by the pianists and the characteristics of the sensors used in this paper, piano playing evaluation metrics were developed that can assess the piano playing process from multiple dimensions.

The three dimensions that affect piano playing are sound, movement and emotion. Based on these three dimensions, the following five piano metrics were introduced:

Pitch: Pitch is the most important indicator in all performance evaluation.

Duration: Whether the duration of each note is accurate or not will directly affect the rhythm of the whole performance.

Fingering: It is very important to play with the correct fingers, which will help them develop good habits.

Depth: the depth of the keys will vary within the same phrase.

Muscle stiffness: this indicator to reflect the emotional state of the user.

III. B. Behavioral understanding phase design

III. B. 1) Microphone Modal Data Processing and Analysis

The most important factors affecting the performance of piano playing are the acoustic dimensions pitch and duration. For the hardware configuration of the system in this paper, it is a natural choice to use a microphone to recognize these two metrics.

(1) Pitch Recognition

In this paper, short-time energy difference features are used to determine the start and end points of a note, and thus the pitch length of the note. Specifically, the short-time energy of the original audio signal is first calculated using a sliding window, i.e:

$$E_i = \sum_{n=0}^{L-1} | x(n) |^2 \tag{20}$$

where $x(n)$ denotes the amplitude of the n rd point in the i nd frame of the signal, and L denotes the window length. The value of the window length is related to the sampling frequency F_s.

(2) Pitch Recognition

Consider using Depth Separable Convolution (DSC) to reduce the amount of computation.

Depth separable convolution differs from standard convolution in that it divides the convolution process into two steps, depth convolution and point convolution, which not only allows for effective extraction of features from the image data, but also reduces the number of parameters and the amount of computation for convolution.

The number of convolution kernel channels used in the calculation of standard convolution is consistent with the input data. The number of convolution kernels is the number of output channels, and the convolution process requires sliding the convolution kernel on the input matrix, multiplying and summing the elements where the convolution kernel slides to, and obtaining the corresponding position data of the output matrix. The weights of the convolution kernel are shared over the entire input matrix, and the computational complexity is relatively high.

In contrast, in the depth-separable convolutional structure, the input data is first subjected to deep convolutional computation. In this process, each convolution kernel has only one channel, the number of convolution kernels is the same as the number of input channels, and the number of output channels is also the same as the number of input channels. An independent convolution operation is performed for each channel of the input data to generate the results of deep convolutional computation for multiple channels. It is worth noting that the operation for each channel involves only the multiply-add operation within the current channel, without the need to perform a cumulative sum operation on multiple channels of data. Next, linear joining is performed by using dot convolution, a process that is equivalent to standard convolution with the size of the convolution kernel fixed to 1×1.

Compared to standard convolution, deeply separable convolution reduces the amount of computation and the number of parameters of the neural network, while increasing the computational efficiency while maintaining the performance of the model. This structure is widely used in resource-constrained scenarios such as mobile devices and embedded systems to accelerate the inference process of deep learning models.

Let the number of channels, height and width (CHW) size of input data be $C_{in} \times N \times N$, the size of convolution kernel of standard convolution and depth convolution be $C_{in} \times K \times K$, the number of output channels be C_{out}, and the step size be 1, then the number of standard convolution parameters and the computation amount are respectively:

$$P_{SC} = C_{in} \times K \times K \times C_{out} \tag{21}$$

$$C_{SC} = C_{in} \times N \times N \times K \times K \times C_{out} \tag{22}$$

And the number of parameters and the amount of computation required to produce a depth-separable convolution with the same size output are:

$$P_{DSC} = C_{in} \times K \times K + C_{in} \times C_{out} \tag{23}$$

$$C_{DSC} = C_{in} \times N \times N \times K \times K + C_{in} \times N \times N \times C_{out} \tag{24}$$

The parametric and computational ratios of the depth-separable convolution to the standard convolution can be computed as respectively:

$$R_P = \frac{P_{DSC}}{P_{SC}} = \frac{1}{C_{out}} + \frac{1}{K^2} \tag{25}$$

$$R_C = \frac{C_{DSC}}{C_{SC}} = \frac{1}{C_{out}} + \frac{1}{K^2} \tag{26}$$

Therefore, from the ratio R_P, R_C, it can be seen that for the same input and output sizes, the number of parameters required and the amount of computation can be significantly reduced by using the Deep Separable Quotient Convolution as compared to the Standard Convolution.

In addition to this, the SE Block attention mechanism is used to improve the feature representation of the model. DSC is used to extract features and then SE Block is used to learn the importance distribution of features. Finally, classification is performed by two fully connected layers. Up to this point, the recognition of pitch and tone length is achieved based on microphone modal data.

III. B. 2) Camera Modal Data Processing and Analysis

Camera modal data preprocessing: the signals from the microphone are utilized to guide the extraction of key frames from the video data.

Camera modal neural network: after data preprocessing, two tasks need to be accomplished: recognizing two metrics: fingering and depth. Utilizing the commonality between the tasks, a multi-task network is designed to learn the two metrics simultaneously.

The camera modal data processing model includes a network skeleton based on DSC, SE Block attention mechanism, and a multi-task learning method based on uncertain parameters.

Up to this point, the camera signal-based action dimension recognition module is successfully constructed, which is an important part of this paper's system to understand the user's piano playing process.

III. B. 3) IMU Modal Data Processing and Analysis

IMU Modal Data Preprocessing: The IMU sensors in this paper's system consist of a 3-axis accelerometer and a gyroscope, which have a sample rate of 100 Hz. Similar to the camera modality, the microphone-based data is first cropped and labeled when the user presses the piano keys through a 1-second time window. Then, a Savitzky-Golay smoothing filter is used to reduce the random noise while preserving the shape of the waveform. In this way, normalized and denoised IMU data was obtained. Finally, the time-frequency domain features in the data are extracted using a short-time Fourier transform algorithm to obtain a spectrogram as input to the subsequent neural network.

IMU modal neural network: after obtaining the spectrogram, a lightweight neural network was constructed to identify muscle stiffness. The metric was defined into two categories, tight and relaxed, thus the identification of this metric is a classification problem.

III. C. Design of the performance scoring phase

“Scoring” refers to the evaluation of the correctness or incorrectness of the behavior during the playing process in order to judge the quality of the overall playing process, which is essentially a comparison of two sets of data. Firstly, all the metrics corresponding to the standard performances according to the repertoire are entered into the system as Groundtruth, and then, by comparing the results of the comprehension phase of the system with the Groundtruth, the scoring of the user's performance process is realized. For this purpose, a scoring method based on the shortest edit distance algorithm is designed.

Scoring process: the two pitch sequences of the recognized results are first aligned, and then the other metrics are compared. For example, for a simple melody S_1: “$Do\ Re\ Mi\ Fa\ So$”, the user played S_2: “$Do\ \underline{Si}\ Mi\ Fa\ So\ \underline{La}$”.

By aligning the note sequences, it can be found that the user played the wrong note " $\underline{Si}$ " and played an extra note " $\underline{La}$ ".

The alignment of S_1 and S_2 is a dynamic programming (DP) problem that can be solved by the shortest edit distance algorithm. The state transfer table is saved during the shortest edit distance algorithm to obtain the note correspondence between S_1 and S_2. After that, each indicator corresponding to the correct note is compared to determine whether it is correct or not. For example, if the " Do " in S_2 is a correctly played note, the length, fingering, and depth of the corresponding note will be further determined.

III. D. Experimental evaluation

III. D. 1) Deep Neural Network Based Gesture Recognition

The basic gestures of piano playing include finger passing, finger passing, finger crossing, finger expanding and finger shrinking. Modern piano performance techniques include percussion techniques, combined with string techniques, etc. In Cowell's The Tides of Manaunaun, the tone clusters are performed with the forearm. In Cowell's The Tides of Manaunaun, the tone cluster technique requires the use of the forearm, including chromatic scales, white and black keys within an octave, arpeggio patterns, and a set of sounds played simultaneously. The Tiger is the most extreme of the clusters, requiring the palm of the hand, both forearms together, and the fist, etc. In most cases, the natural weight of the arm itself is sufficient, and there is no need to add extra strength. There are many other techniques for playing clusters, such as forearm clusters, hammer-ons on the strings, slides on the strings, and silent pressing of chords. Combining piano with strings is a complete reversal of keyboard playing, shifting parts to the strings inside. Piano strings are handled in piano music in a variety of ways, including: scraping or plucking, pulling techniques on the strings, mute techniques on the strings, and tonal piano concepts.

Piano playing gestures change according to the repertoire, style of playing, etc. In this paper, standard piano playing is used here as a benchmark for judging, focusing on analyzing the finger movements of each of the four basic gestures of the performer.

The image data in the piano playing gesture dataset is allocated in the ratio of 9:1 to obtain the training set and test set. The assigned datasets for the training set and test set are shown in Table 2.

Among them, the training set is used to instruct the neural network to learn the ability to classify and recognize piano playing gestures, while the test set is used to evaluate the generalization ability of the deep convolutional neural network.

Table 2: The distribution of the training set and the test set is set

Subset	Training set	Test set
Quantity (sheet)	4251	569

Experimental environment: utilizing GPU as an experimental platform, the server is configured with the following: processor Intel Core i9-9900k with a main frequency of 3.6GHz and four graphics cards Ge Force RTX 2080Ti. The neural network was built using the Keras framework in the Python programming language environment to complete the experiments in this chapter.

In this chapter, VGG19, ResNet50, and InceptionV3 image recognition networks were used for experiments respectively, and the comprehensive piano playing hand shape recognition accuracy is shown in Table 3.

The neural network model based on DSC's network skeleton fused with SE Block attention mechanism proposed in this paper has a comprehensive recognition rate of 93.54% for piano playing hand gestures (down-finger, through-finger, across-finger, expanding and contracting finger).

The experimental results show that the design of the solution for the intelligent recognition of piano playing hand shape is feasible, based on the classification of piano hand shape, the deep neural network image recognition method can effectively realize the intelligent recognition of piano playing hand shape. The neural network model based on DSC network skeleton fusion SE Block attention mechanism can meet the needs of the piano playing intelligent evaluation system designed in this paper. After experimental processing, the algorithm also meets the requirements of microphone modal data processing, and can effectively recognize the pitch and duration of the piano performance.

Table 3: Comprehensive piano playing hand recognition accuracy

Piano gesture	Finger	VGG19	ResNet50	InceptionV3	Chapter algorithm
Cis finger	Thumb	0.9235	0.9124	0.9425	0.9651
	Index finger	0.8936	0.8253	0.8693	0.9224
	Middle finger	0.9021	0.8563	0.8747	0.9056
	Ring finger	0.8693	0.8609	0.8573	0.9201
	Little finger	0.8925	0.8758	0.8639	0.9127
Cross fingers	Thumb	0.8565	0.8321	0.8758	0.9213
	Index finger	0.8251	0.8024	0.8969	0.9042
	Middle finger	0.8063	0.8059	0.8724	0.9125
	Ring finger	0.8124	0.8213	0.8563	0.9263
	Little finger	0.7893	0.8346	0.8692	0.9546
Weave fingers	Thumb	0.7505	0.7872	0.8621	0.9631
	Index finger	0.7681	0.7809	0.8463	0.9514
	Middle finger	0.7758	0.7993	0.8365	0.9221
	Ring finger	0.7802	0.7954	0.8427	0.9335
	Little finger	0.7962	0.8031	0.8596	0.9245
Expansion finger and shrinkage finger	Thumb	0.7036	0.7253	0.8110	0.9364
	Index finger	0.7103	0.7365	0.8124	0.9427
	Middle finger	0.7215	0.7403	0.8362	0.9536
	Ring finger	0.7356	0.7509	0.8127	0.9708
	Little finger	0.7448	0.7359	0.8163	0.9652
Comprehensive recognition accuracy		0.8029	0.8041	0.8557	0.9354

III. D. 2) Fingerprint Audio Comparison Experiments

For the fingering audio, its frequency domain information is obtained by Fourier transform. And compared with the corresponding standard tones. Taking chromatic scale 4 (fa) as an example, its time domain waveform is shown in Fig. 7. The Amp value of the played chromatic scale 4 (fa) is within [-1,1], which is consistent with the piano chromatic scale 4 (fa) audio.

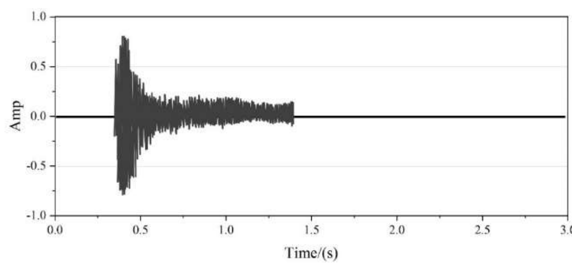

Figure 7: Semi-tone order 4(fa) time domain results

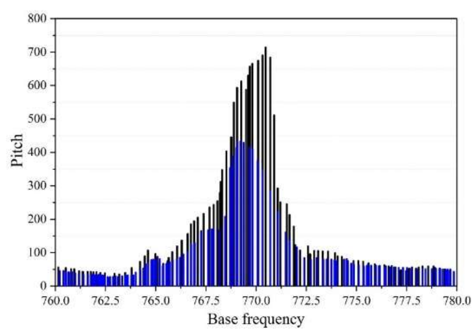

Figure 8: Semi-tone order 4(fa) frequency domain results

The frequency domain result of chromatic scale 4 (fa) is shown in Figure 8, which clearly shows its fundamental frequency information. The frequency value obtained is consistent with the standard tone of 780 Hz for the middle tone fa in D major, this chromatic scale is in correct pitch, which indicates that the left hand presses the strings with reasonable strength, so the evaluation result is that the audio of this fingering is correct.

Intonation is crucial for learning a musical instrument, and the use of pitch comparison for audio assessment is simple and effective. It has higher accuracy compared to video evaluation. Thus the neural network model based piano playing evaluation system designed here in this paper is able to score effectively.

IV. Conclusion

This paper calculates the similarity between high-dimensional data and proposes a serialized recommendation algorithm for collaborative filtering of high-dimensional data using deep neural networks. Piano playing judging criteria are proposed from the piano playing dimension, and a piano playing behavior stage recognition algorithm is formulated based on deep separable convolutional networks.

(1) Filtered multidimensional dataset to compare the data recommendation algorithm based on improved deep neural network with the classical sequence recommendation algorithm. On the news recommendation dataset, the click-through rates of the HitRate@50 and HitRate@150 indicators based on the improved deep neural network are 0.151% and 0.279%, respectively, which are significantly higher than those of other models. On the MovieLens 1M dataset, the click-through rate of the HitRate@50 index based on the improved deep neural network data recommendation algorithm is 60.59%, which is 37.74% higher than that of the traditional FM model.

(2) The neural network model formed by a deep separable convolutional neural network fused with the SE Block attention mechanism has a recognition rate of 93.54% in the camera modal data processing of the piano playing evaluation system, and at the same time can effectively evaluate the fingering audio and objectively deal with the multi-dimensional data of piano playing gestures.

Funding

This study is supported by “Guangdong Undergraduate Universities Online Open Course Steering Committee, China” (Grant No.2022ZXKC301).

References

[1] Chang, H., & Panicker, C. M. V. (2024). Exploring Trends in Piano Education and Music Literacy Across China From 20th and 21st Century. Sciences of Conservation and Archaeology, 36(3), 462-472.
[2] Guobin, Z., Suttachitt, N., Charoensloong, T., & Daoruang, K. (2025). The Development of Online Lessons for Promoting Piano Playing Skills for Non-piano Playing Background Students. Journal of Ecohumanism, 4(1), 3671-3696.
[3] Mitchell, N. (2017). Beginning and intermediate piano students' experiences participating in evaluative performances. International Journal of Music Education, 35(4), 515-525.
[4] Cui, Y. (2023). Vocal Music Performance Evaluation System Based on Neural Network and Its Application in Piano Teaching. Revista Ibérica de Sistemas e Tecnologias de Informação, (E55), 451-464.
[5] Phanichraksaphong, V., & Tsai, W. H. (2021). Automatic evaluation of piano performances for STEAM education. Applied sciences, 11(24), 11783.
[6] Chen, Y., & Zheng, N. (2021). Ai based research on exploration and innovation of development direction of piano performance teaching in university. Journal of Intelligent & Fuzzy Systems, 40(2), 3681-3687.
[7] Jiang, Y. (2023, November). Expert and Novice Evaluations of Piano Performances: Criteria for Computer-Aided Feedback. In ISMIR (pp. 367-374).
[8] Hamond, L. F., Welch, G., & Himonides, E. (2019). The pedagogical use of visual feedback for enhancing dynamics in higher education piano learning and performance. Opus, 25(3), 581-601.
[9] Tabuena, A. C. (2019). Effectiveness of classroom assessment techniques in improving performance of students in music and piano. Global Researchers Journal, 6(1), 68-78.
[10] Li, W. (2022). Analysis of piano performance characteristics by deep learning and artificial intelligence and its application in piano teaching. Frontiers in Psychology, 12, 751406.
[11] Kim, H., Ramoneda, P., Miron, M., & Serra, X. (2022). An Overview of Automatic Piano Performance Assessment within the Music Education Context. CSEDU (1), 465-474.
[12] D. Dhinakaran,L. Srinivasan,S. Edwin Raja,K. Valarmathi & M. Gomathy Nayagam. (2025). Synergistic feature selection and distributed classification framework for high-dimensional medical data analysis. MethodsX,14,103219-103219.
[13] Yi Wang,Wenshan Li,Tao Li & Hao Tian. (2025). Single-stage filter-based local feature selection using an immune algorithm for high-dimensional microarray data. Applied Soft Computing,172,112895-112895.
[14] Hyeonbin Moon,Kundo Park,Jaemin Lee,Donggi Lee & Seunghwa Ryu. (2025). DNN-based inverse design of line heating patterns for automated plate forming in shipbuilding using multi-start convex optimization. Extreme Mechanics Letters,76,102313-102313.