

A Conv-TCN Model for Extracting Dynamic Rhythm Paths in Expressive Piano Performances

Xiaolan Qiu¹, Buwen Hou¹, Xin Wang²

¹College of Preschool Education (Music College), Lianyungang Normal University, Lianyungang 222100, Jiangsu, China

²The Logistics Management Office, Jiangsu Lianyungang Vocational and Technical College of Industry and Trade, Lianyungang 222100, Jiangsu, China

Corresponding author: Xin Wang (e-mail: 18552233885@163.com)

ABSTRACT Accurate characterization of expressive rhythm evolution remains a challenging task because conventional methods primarily rely on static descriptors and fail to capture long-range temporal dependencies. This study proposes a Conv-TCN framework for extracting dynamic rhythm paths from piano performance data by integrating one-dimensional convolutional neural networks with temporal convolutional networks. Temporal features are first extracted from MIDI sequences and encoded to represent local rhythmic structures, while dilated temporal convolutions model long-term dependencies and gradual expressive variations. A multi-scale fusion strategy further combines rhythmic information across different temporal resolutions to generate continuous and interpretable rhythm trajectories. Experimental results demonstrate that the proposed method achieves an MSE of 0.032, smoothness of 0.020, KL divergence of 0.052, and similarity of 0.83 under noisy conditions, outperforming several representative sequence modeling approaches. Beyond music performance analysis, the proposed temporal feature extraction and dynamic trajectory modeling strategy provides a methodological reference for processing complex non-stationary signals, with potential value for electromagnetic signal characterization, antenna measurement data analysis, and propagation-related time-series interpretation.

INDEX TERMS dynamic rhythm modeling, temporal convolutional network, multi-scale feature fusion, temporal signal analysis, electromagnetic signal processing

I. INTRODUCTION

THE concept of rhythm is extensive in its interaction with composition, underscoring the meaningfulness of temporal aspects in music and affirming the possibility of expressive qualities in performance. Exposure to piano playing demonstrates how music and technical skill manifest through deliberate rhythmic articulation—a sustained temporality of precision akin to the persistent, ongoing creations in manufacturing. For example, the rhythmic application in weaving fabric or in the complexity of design relies on accurately timed, repeated feeding of yarn and shuttle. Temporal behaviors such as tempo gradients, delay, and acceleration are fundamental tools for dynamic expression in music [1,2]. Similarly, composers transform mechanical timing through rhythmic processes (e.g., rubato, ritardando, accelerando), enabling greater human expression in piano compositions [3,4]. This concept of dynamically controlled trajectories is also relevant for industrial control systems in the industry, where maintaining a soft, non-rigid flow of materials is essential to prevent fabric tearing or warping while maximizing throughput.

II. RELATED WORK

Conventional rhythm analysis techniques typically employ time window averaging, fixed-length segment extraction, and assess only the general rhythmic level in performances (e.g., average tempo and tempo stability). While these methods offer practical utility in tasks such as automatic music scoring and basic performance analysis, their expressiveness is limited in modeling performance styles and recognizing individual expressive qualities. Other research has focused on rhythm type classification, beat prediction, and time-series modeling using reinforcement learning algorithms to describe temporal and spatial changes in rhythmic patterns [5,6]. Multimodal emotion analysis models based on auditory and visual features have been introduced to explore the relationship between music, emotion, and scene [7]. Recurrent neural networks (RNNs) and long short-term memory (LSTM) variants have been utilized to address long-term dependencies in rhythm sequences, improving prediction accuracy and style transfer capabilities [8,9]. However, these approaches primarily focus on prediction or classification, lacking the ability to model and visualize the rhythm change trajectory itself, thereby failing to describe the natural evolution of rhythm during performance [10,11].

Aggregated metrics such as average tempo deviation or fixed-window timing variability reduce rich temporal dynamics to scalar values or isolated segments, omitting the continuous evolution of rhythm. In contrast, the proposed convolutional-temporal convolutional network (Conv-TCN) model provides a fundamentally dynamic framework for rhythm analysis. Rather than relying on static summary statistics, it represents the rhythmic path as a continuous, high-resolution time series. The one-dimensional convolutional neural network (1D-CNN) component captures local, short-term rhythmic structures, while the dilated TCN models long-range temporal dependencies, enabling the tracking of gradual expressive variations such as rubato and ritardando across entire musical phrases. Furthermore, the multi-scale fusion mechanism integrates local and global features into interpretable, evolving trajectories, revealing how rhythmic tension and tempo develop progressively over time. By preserving the full temporal structure of expressive timing, this approach overcomes the limitations of conventional static representations.

Meanwhile, music visualization has emerged as an important approach for understanding the multidimensional structure of musical information and diverse performance practices. For example, Georges P. et al. employed network-based representations and support vector machines to cluster compositional styles, demonstrating new possibilities for music data analysis [12]. Liu F. et al. proposed a self-supervised framework for music-driven action synchronization, highlighting the potential of cross-modal modeling in performance analysis [13]. In the domain of piano performance modeling, Jia H. emphasized the intrinsic relationship between technical proficiency and musical expression, proposing a theoretical framework for performance style modeling [14]. Similarly, Peng L. utilized a lightweight convolutional network for intonation training, validating the effectiveness of deep learning in individualized performance assessment [15]. In addition, Ghatas Y. et al. introduced a hybrid deep learning approach estimating the difficulty of symbolic piano music, further expanding the applicability of intelligent systems in performance assistance [16]. This paper addresses the inadequacies of existing rhythm modeling by introducing quantitative markers for rhythm path dynamics and contrasting static and temporal models in terms of error performance, demonstrating the superiority of dynamic modeling. The Conv-TCN model leverages temporal characteristics and multi-scale dilated convolutions for comprehensive rhythm modeling, evaluated with real-world performance data and industry tools, and compared to style transfer and multimodal modeling applications. This research improves the ability to express timing in rhythmic modeling, which also functions as a modeling tool for situating performance styles and modelling strategies around rhythm control. This baseline, for example, will also prove useful for my future aspirations around performance style identification, or the invention of new performances. Likewise, the enriched temporal modeling could help sup-

port a dynamic control approach that would provide more customized constructions as the machines would be able to respond dynamically, in real-time, to different batches of material, or, complex, customized fabric designs.

III. MODEL CONSTRUCTION

A. INPUT DATA PROCESSING: RHYTHM TIMING FEATURE EXTRACTION

Piano MIDI performance data were used to systematically derive rhythmic timing features, preserving the organized temporal characteristics of the input data [17,18]. Each MIDI file was parsed to extract note event parameters—start time, end time, pitch, and velocity. Rhythm-focused features included note start and end times, inter-onset intervals (IOI), beats per minute (BPM), beat position, and accent intensity [19,20]. BPM was calculated directly from IOIs of consecutive note onsets, providing an intrinsic measurement of instantaneous tempo. All features were linearly normalized to the range [0, 1] and resampled to a uniform sampling rate of 20 Hz (one time point every 50 ms) using linear interpolation to fill missing values. The resulting multi-dimensional feature sequences were concatenated along the time axis, forming a unique input tensor $[N, T, F]$ representing the relationship across time with $[X, Y, Z]$, where N is the sample number, T is the number of time steps, and F is the feature dimension ($F = 5$).

Rigorous data quality control was implemented: performance samples shorter than 30 seconds were excluded; median filtering was applied to IOI data to remove abnormal intervals and sampling noise; extreme BPM values (> 300 or < 30) were pruned based on statistical thresholds; beat position errors were corrected by comparing score beat information and metronome marks; Min-Max normalization balanced feature distribution and limited channel bias.

Figure 1 illustrates the extraction and processing of rhythm timing features. The horizontal axis represents playing time (seconds), and the vertical axis represents normalized rhythmic feature values. IOI, BPM/200, and beat position demonstrate temporal changes. BPM shows slower periodic fluctuations corresponding to long-term rubato; IOI exhibits faster local rhythm fluctuations; beat position maintains a cyclic distribution between 0 and 1, reflecting rhythmic periodicity. Smoothing yields more continuous curves, suppressing local jitter and clarifying dynamic trends. Histograms show IOI is concentrated between 0 and 0.4, BPM/200 fluctuates between 0.2 and 0.55 (approximately 40–110 BPM), and accent intensity is widely distributed, validating free rhythmic emphasis in piano performance.

B. LOCAL RHYTHM STRUCTURE ENCODING: CONVOLUTIONAL LAYER FOR SHORT-TERM RHYTHM MORPHOLOGY

A 1D-CNN was employed to encode local structures in the multi-dimensional rhythm feature sequence, capturing short-term changes in rhythm morphology. This module extracts key short rhythm characteristics—rhythm type, repetitive

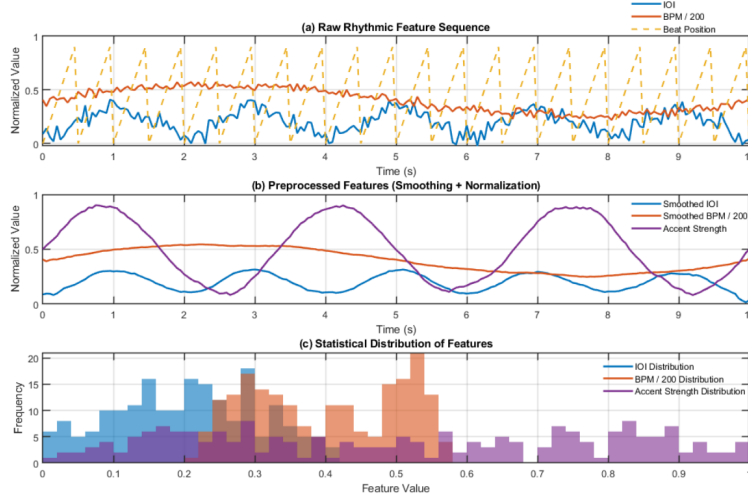


FIGURE 1. Piano Performance Rhythm Timing Feature Extraction and Distribution Analysis: (a). Original rhythm feature timing curve; (b). Preprocessed rhythm feature smoothing curve; Rhythm feature distribution histogram

patterns, local strong and weak structures—via sliding window convolution operations, providing rich local dynamic information for temporal modeling [21].

The input consists of a three-dimensional tensor of dimensionality $X \in \mathbb{R}^{N \times T \times F}$, where N represents the sample batch size, T is the number of time steps, and $F = 5$ refers to the rhythmic feature dimensions for each time step (note onsets and endpoints, IOI, BPM, beat position, accent intensity). The convolutional layer uses a kernel width of 7 time steps (~350 ms), suitable for short-term variations in piano rhythm. For each convolution kernel, the output feature at time step t is calculated as follows:

$$y_t^{(k)} = b^{(k)} + \sum_{f=1}^F \sum_{i=0}^6 w_{i,f}^{(k)} \cdot x_{t+i-[7/2],f} \quad (1)$$

Here, $y_t^{(k)}$ represents the output of the k -th convolution kernel at time step t , $w_{i,f}^{(k)}$ is the weight of the k -th convolution kernel corresponding to the i -th time offset and the f -th feature channel, and $b^{(k)}$ is the bias term. To effectively cover every time point separate from edge effects, the input data is symmetrically padded. Therefore, convolution performs a weighted summation across all channels and time steps to extract local temporal features. The output of convolution is passed through a nonlinear activation function. In this paper, the ReLU function is used as given in Equation (2):

$$z_t^{(k)} = \max(0, y_t^{(k)}) \quad (2)$$

ReLU activation gives the model nonlinear modeling capabilities, enabling it to effectively represent complex local rhythm changes. To improve training stability and accelerate convergence, batch normalization is then used to standardize the mean and variance of the activation values, as shown in Equation (3):

$$\hat{z}_t^{(k)} = \frac{z_t^{(k)} - \mu^{(k)}}{\sqrt{(\sigma^{(k)})^2 + \epsilon}} \quad (3)$$

$\mu^{(k)}$ and $\sigma^{(k)}$ are the mean and standard deviation of the k -th convolution channel in the batch, respectively, and ϵ is a small constant to prevent division by zero. Batch normalization standardizes activation values, alleviating internal covariate shift and ensuring stable gradient propagation during training. Three stacked 1D convolutional layers are used, with 64, 128, and 256 kernels respectively. Kernel width and stride are maintained at 7 and 1, with symmetric padding. Residual connections between layers facilitate information flow and gradient transmission. The residual connection is defined as Equation (4):

$$H^{(l+1)} = F(H^{(l)}) + H^{(l)} \quad (4)$$

Here, $H^{(l)}$ and $H^{(l+1)}$ represent the output features of the l and $(l+1)$ layers, respectively, and F represents the convolutional transformation and subsequent activation and normalization operations of that layer. This design alleviates the vanishing gradient problem in deep networks and ensures the effective transmission of local rhythm information.

Figure 2 depicts the iterative processing of input features, highlighting the role of residual connections in encoding local rhythmic structures. The x-axis represents time (s), and the y-axis denotes features amplitude, corresponding to the strength of the rhythm signal. In Figure 2(a), the gray line illustrates the original input signal, which consists of a mixture of low- and high-frequency components with superimposed random noise, reflecting the inherent complexity of expressive piano performance. The blue, orange, and purple lines represent the feature outputs from three convolutional layers, showing a hierarchical progression from coarse, smoothed representations to more detailed rhythmic features.

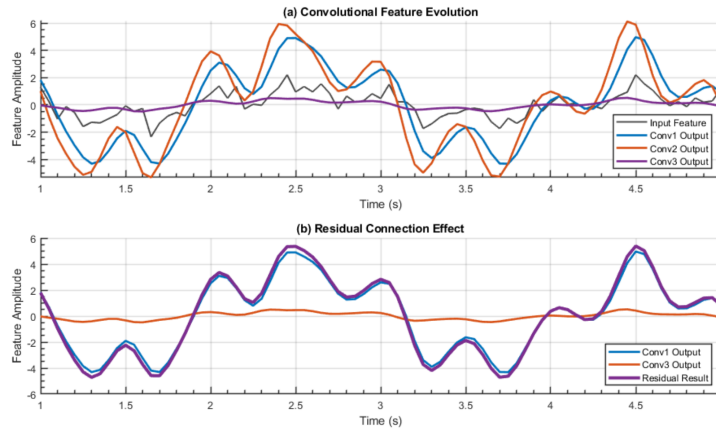


FIGURE 2. Local Rhythm Structure Encoding: (a) Feature evolution through convolutions; 2(b) Residual connection impact

Over successive layers, the convolutional kernels progressively refine the signal representation, enhancing relevant rhythmic patterns while suppressing noise. In Figure 2(b), a residual connection is introduced between the first and third convolutional layers. This design preserves global signal continuity while enriching local feature representation, resulting in a more stable and expressive rhythmic texture. Consequently, the model achieves improved encoding of subtle rhythmic variations.

C. TEMPORAL MODELING: TCNS FOR LONG-TERM RHYTHMIC EVOLUTION

1) TCN Architecture and Dilated Convolution Mechanism

TCNs were developed to model long-term dependencies in rhythm feature sequences. The fundamental building block is dilated convolution, which expands the receptive field without significantly increasing learnable parameters or computational complexity. The dilated convolution operation is defined as the convolution of the input sequence $x = \{x_0, x_1, \dots, x_{T-1}\}$ at the time step t as follows:

$$y(t) = \sum_{i=0}^{k-1} w(i) \cdot x_{t-d \cdot i} \quad (5)$$

k is the convolution kernel size, $w(i)$ is the convolution kernel weight, and d represents the dilation factor, which determines the sampling interval. Exponentially increasing the dilation factor (for example, $d = 1, 2, 4, 8, \dots$) layer by layer enables coverage of long time scales, capturing gradual changes and long-term fluctuations such as *ritardando* and *rubato*.

2) Model Configuration and Training Procedures

TCN utilizes causal convolution to prevent future data leakage, ensuring that outputs depend only on current and past inputs. The input is the multi-channel local rhythm feature sequence from the convolutional encoding module. The TCN processes these features layer by layer, producing a rhythm evolution path sequence that preserves long-term

dependencies and temporal dimension, facilitating subsequent multi-scale feature fusion. The Adam optimizer is used for training, with an initial learning rate of 0.0005 and a decay schedule, batch size of 16. The loss function combines mean squared error (MSE) and dynamic time warping (DTW) distance, enhancing sensitivity to subtle rhythm changes.

D. MULTI-SCALE RHYTHM FUSION: INTEGRATING LOCAL AND GLOBAL PATHS

1) Multi-Level Output Design and Skip Connection Mechanism

A multi-level TCN output configuration accounts for multidimensional rhythmic feature variations. Independent output branches after each dilated convolutional layer extract rhythmic representations within each receptive field. Skip connections combine these features, preventing information loss and vanishing gradients, and enabling parallel streaming of rhythms across time scales. The skip connection layer employs time-step-by-time weighted summation to fuse multi-scale feature maps into a unified rhythm path sequence calculated as follows:

$$Y(t) = \sum_{l=1}^L \alpha_l \cdot X_l(t) \quad (6)$$

Here, L represents the number of skipped layers, $X_l(t)$ represents the representation of the output features of layer l at time step t , and the weight coefficient α_l is automatically learned through training to ensure the optimal combination of information at each scale.

Normalization batch-normalizes the output features of each layer along the temporal dimension, ensuring that features at different scales have similar numerical ranges and distributions, facilitating balanced calculations in subsequent weight fusion.

Sliding window smoothing (window length: 5 time steps) suppresses noise and abnormal fluctuations, improving expressiveness and interpretability.

Figure 3 presents the rhythm path features extracted from a piano performance using the proposed multi-scale fusion strategy, along with their time–frequency analysis. In Figure 3(a), the horizontal axis denotes time (s), the left vertical axis represents the fused rhythm output, and the right vertical axis indicates the adaptive weights assigned to short-, medium-, and long-term rhythmic features. At approximately 12.5 s, the short-term weight reaches a peak value of 0.58, while the long-term weight decreases to around 0.02. At this stage, the fused rhythm path exhibits pronounced high-frequency oscillations, effectively capturing fine-grained ornamental details. In contrast, at around 17 s, the long-term component dominates (weight = 0.63), leading to a smoother and more gradually evolving output trajectory.

Figure 3(b) illustrates the continuous wavelet transform (CWT) spectrum of the fused rhythm path, where the x-axis represents time and the y-axis represents frequency (Hz). Color intensity reflects the magnitude of frequency components. The low-frequency components persist throughout the entire duration, indicating the stability of the global rhythmic structure. In contrast, mid-frequency bands exhibit localized intensification within the intervals of 0–5 s, 10–16 s, and 20–25 s, corresponding to mesoscale rhythmic oscillations. Overall, the proposed fusion strategy enables dynamic reweighting of rhythmic information across multiple temporal scales, allowing the model to capture both global structural stability and local expressive variations. This enhances the model's capability for accurately representing complex rhythmic evolution in piano performance.

2) Parameter Configuration and Performance Optimization

Five skip connections correspond to outputs of five TCN layers with dilation factors of 1, 2, 4, 8, and 16, covering a temporal receptive field of 3–97 sample steps (20 samples/sec). The fusion layer has 128 output channels, balancing computation and feature expressiveness. The weight coefficients (α_l) are initialized from a uniform distribution and subsequently optimized via backpropagation during training. This enables the model to adaptively emphasize rhythmic information at relevant temporal scales while suppressing uninformative or redundant features.

The fusion layer applies linear transformations followed by ReLU activation functions to enhance the representation of nonlinear characteristics. To mitigate overfitting, a dropout mechanism is incorporated into the skip-fusion component, with a dropout rate of 0.2 to improve model generalization.

The entire multi-scale fusion module is trained end-to-end together with the preceding TCN encoding layers, enabling dynamic optimization of the global rhythmic path representation.

E. RHYTHM PATH OUTPUT STRUCTURE: NORMALIZATION AND TRAJECTORY RECONSTRUCTION

1) Temporal Normalization

Following the multi-layer TCN modeling and multi-scale fusion process, a rhythmic path sequence of fixed length is obtained, determined by the predefined sampling frequency and time window size. However, due to variations in performance duration and tempo across different piano recordings, a temporal normalization procedure is required to enable consistent comparison across samples.

In this study, time normalization is implemented using linear interpolation to map the original rhythm state sequence onto a unified temporal axis. The normalized sequence length is typically set between 200 and 500 time steps, selected based on the average performance duration and the sampling frequency of the training data. This design balances the preservation of temporal detail with computational efficiency.

Formally, for an original path sequence $P = \{p_1, p_2, \dots, p_N\}$, where $p_i \in \mathbb{R}^d$ represents the rhythm state vector at the i -th time step and N is the original sequence length, a normalized sequence is generated via a linear interpolation function $f : [1, N] \rightarrow [1, T]$, which performs temporal stretching or compression, as defined in Equation (7):

$$P' = \{p'_1, p'_2, \dots, p'_T\}, \quad p'_t = p_{f^{-1}(t)} \quad (7)$$

Here, f^{-1} stands for interpolated inverse mapping, ensuring that the normalized sequence maintains the morphology and dynamics of the original rhythmic path.

2) Rhythm State Vector Construction

The model predicts rhythmic state vectors, including current tempo (BPM), rhythmic tension (rubato), and beat offset. Vectors are initialized with dimension $d=16$ and batch normalized to mean 0 and standard deviation 1, improving stability and smoothing output sequences.

3) Sequence Smoothing

To enhance temporal continuity and the interpretability of the rhythmic path sequence, two smoothing strategies are employed: Gaussian filtering and smoothing regularization. The Gaussian smoothing method applies a one-dimensional Gaussian kernel along the temporal axis to convolve the rhythmic state vector at each time step. The kernel has a standard deviation of $\sigma = 2$ and a window size of 7 time steps.

This operation effectively suppresses short-term noise and abrupt fluctuations in the sequence, as defined in Equation (8):

$$\tilde{p}_t = \sum_{k=-3}^3 w_k p_{t+k}, \quad w_k = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{k^2}{2\sigma^2}\right) \quad (8)$$

This method of moving ensures that the path for rhythm change has a natural and smooth temporal form, removing local instability in training or during the model's prediction.

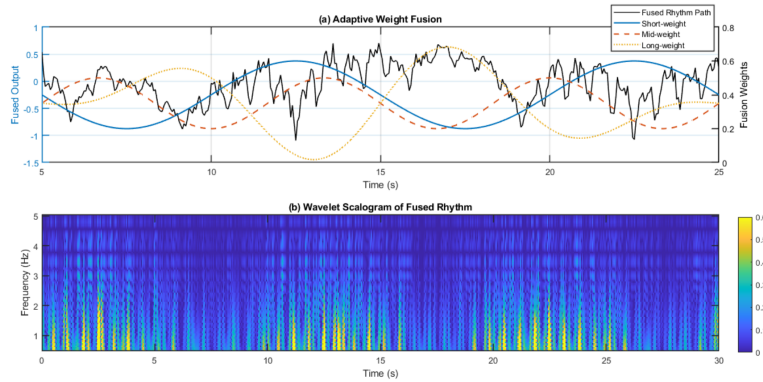


FIGURE 3. Visualization of Multi-Scale Rhythm Feature Fusion: (a). Adaptively fused rhythm path and weight changes; (b). Wavelet time-frequency spectrum of the fused rhythm path

In parallel, during training phase, we add a first-order difference smoothing regularization term, which is given by Equation (9):

$$L_{smooth} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|p_{t+1} - p_t\|_2^2 \quad (9)$$

The regularization promotes a smoother change in the model's output rhythmic paths, which reduces abrupt and nonsensical changes, and increases the temporal consistency and logicity of the trajectory. The regularized and smoothed pattern of rhythmic paths is returned as output, in a format that is structured as a time series, and ready for analysis and application. The output quote states a full representation of your rhythmic characteristics dynamically changing in a piano performance, complete with changes in tempo and small adjustments to beat offset, as well as time-dependent changes to rhythmic tension. The model adopts a hierarchical framework, as shown in Figure 4:

The model employs a hierarchical framework: feature extraction from MIDI data, short-term encoding via 1D-CNN with residual connections, long-term modeling with dilated causal TCN, weighted multi-scale fusion, and output normalization and smoothing. This facilitates end-to-end dynamic rhythm path modeling from raw performance input.

IV. EVALUATION

A. DATASET DESCRIPTION

The evaluation utilizes MIDI performance data from 10 professional pianists across multiple genres (classical, romantic, modern), comprising over 100 pieces totaling more than 12 hours [22]. All MIDI data underwent timing correction, note alignment, and velocity normalization. Expressive rhythmic offsets were annotated as reference standards. The dataset includes diverse playing speeds and techniques, enabling robust model evaluation. Training and test sets were split to ensure objective, generalizable assessments.

B. PHASE TRAJECTORY AND CURVATURE ANALYSIS

To characterize the functional dynamics of nonlinear rhythmic evolution in piano performance, phase-space trajectory

analysis is employed in this study. The three-dimensional phase-space representation in Figure 5(a) and the spatiotemporal curvature field shown in Figure 5(b) provide a differential geometry-based quantification of the dynamic deformation characteristics of rhythmic paths.

Figure 5(a) illustrates the three-dimensional trajectory of the rhythmic path in phase space. The axes correspond to the current state $xx(t)$, the one-step lagged state $xx(t + \Delta t)$, and the two-step lagged state $xx(t + 2\Delta t)$. The resulting trajectory exhibits a spiral-like structure, with color encoding the curvature magnitude (from low in blue to high in red). Regions with high curvature (red clusters) coincide with turning points in the trajectory, corresponding to expressive timing variations such as rubato, and indicating abrupt changes in rhythmic tension.

Figure 5(b) presents the spatiotemporal curvature field of the rhythmic path. The horizontal axis represents normalized time (0–1), the vertical axis denotes the rhythmic state value (–1 to 1), and the color intensity reflects curvature magnitude. The black curve projected onto the surface indicates the actual trajectory of the rhythm path. A prominent high-curvature region is observed in the interval 0.6–1.0, corresponding to the acceleration phase following a ritardando. This pattern reveals localized nonlinear transitions in rhythmic structure, reflecting expressive temporal modulation in performance.

Overall, Figure 5 provides a quantitative characterization of rhythmic path dynamics using differential geometry, demonstrating the inherently nonlinear evolution of micro-timing control in piano performance.

C. FITTING ACCURACY AND SMOOTHNESS

Model fitting accuracy is evaluated using the MSE, which quantifies the discrepancy between the predicted rhythmic path and the manually annotated reference trajectory by averaging the squared errors at corresponding time points. Lower MSE values indicate that the predicted rhythmic path more accurately reproduces the true temporal dynamics of the performance.

Rhythm path smoothness is measured using the standard deviation of the first-order differences of the trajectory. This

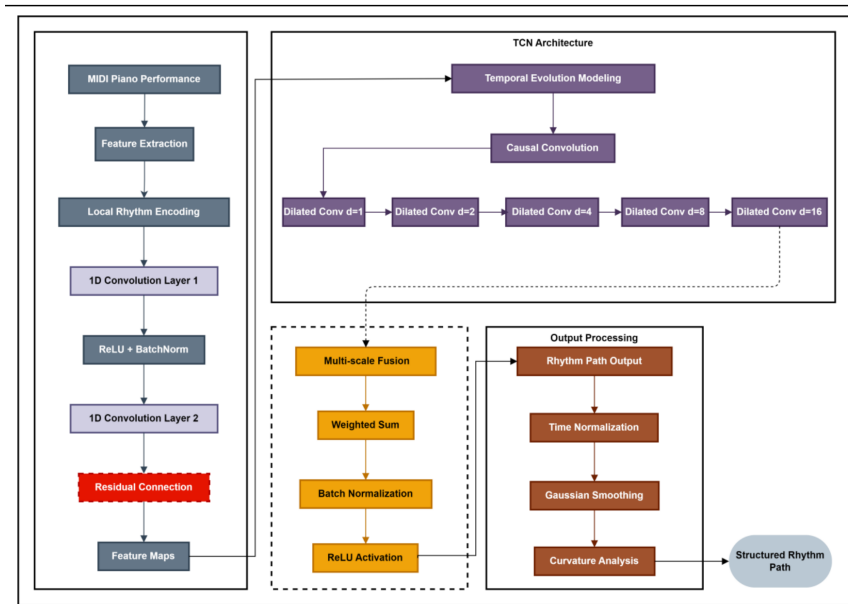


FIGURE 4. Hierarchical Processing Architecture of the Rhythm Path Extraction Model

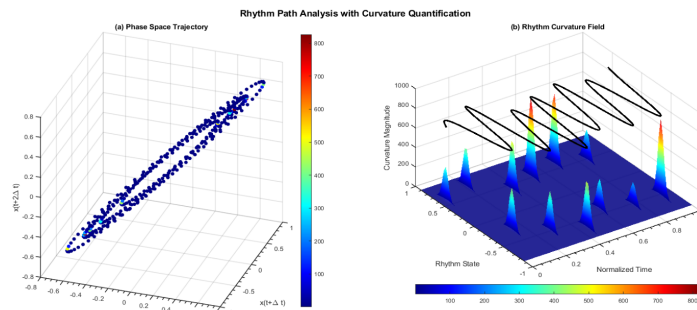


FIGURE 5. Phase Trajectory and Curvature Field Analysis of the Rhythm Path: (a). Phase Space Trajectory; (b). Curvature Field

metric reflects the continuity of rhythmic evolution and the stability of temporal variations. A lower value indicates fewer abrupt fluctuations or noise artifacts in the generated trajectory, thereby ensuring a more natural and coherent representation of rhythmic progression.

The proposed Conv-TCN model is compared with several representative sequence modeling approaches, including LSTM, Transformer, TCN, 1D-CNN, Hidden Markov Model (HMM), Support Vector Regression (SVR), Random Forest, Gated Recurrent Unit (GRU), and WaveNet.

Figure 6 provides a comprehensive comparison of rhythm path modeling performance in terms of fitting accuracy and trajectory smoothness. The horizontal axis represents the evaluation metric—MSE in Figure 6(a) and the standard deviation of the first-order differences in Figure 6(b)—while the vertical axis lists the evaluated models.

As shown in Figure 6(a), the proposed Conv-TCN model achieves the lowest MSE (0.032), indicating superior fitting accuracy compared with mainstream sequence models such as LSTM (0.078) and Transformer (0.065), as well as traditional machine learning approaches including HMM

(0.121) and SVR (0.145). This result demonstrates the model's stronger capability to capture the temporal dynamics of rhythmic evolution. Figure 6(b) further shows that the Conv-TCN model achieves the lowest smoothness indicator (0.020), indicating the most stable and coherent rhythmic trajectories among the compared methods. In contrast, models such as SVR, HMM, and Random Forest exhibit smoothness values exceeding 0.05, reflecting more pronounced trajectory fluctuations and reduced temporal consistency.

Overall, these results confirm that the proposed Conv-TCN framework provides both higher predictive accuracy and smoother rhythmic path representations, highlighting its effectiveness for modeling complex temporal structures in expressive piano performance.

D. COMPARATIVE ANALYSIS OF PERFORMANCE STYLES

DTW distance quantifies rhythmic process differences among pianists, facilitating style comparison. Larger DTW distances indicate greater stylistic divergence. The analysis

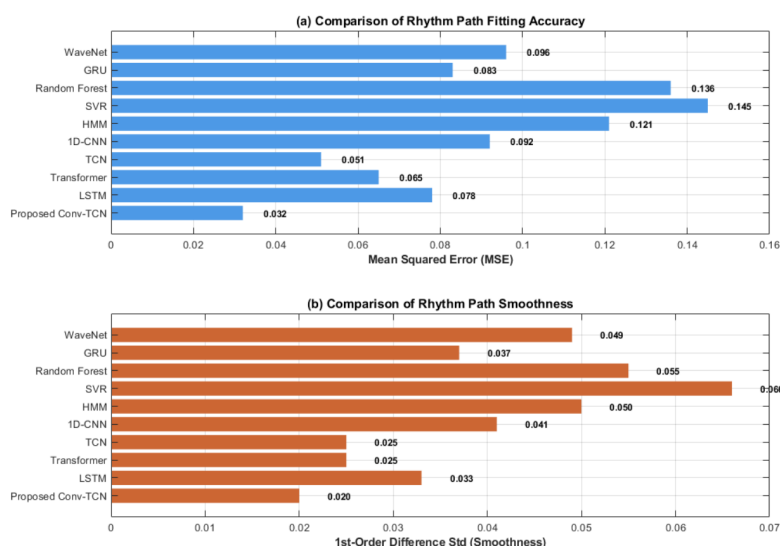


FIGURE 6. Comparison of Rhythm Path Fitting Accuracy and Smoothness of Different Models: (a). Rhythm Path Fitting Accuracy Comparison; (b). Rhythm Path Smoothness Comparison

enables in-depth evaluation of rhythmic expression, identifying personalized evolution pathways.

Figure 7(a) presents the rhythmic style similarity matrix computed using DTW distances among 10 pianists. Both the horizontal and vertical axes correspond to individual pianists, while the color scale represents the degree of similarity, with lower values indicating greater similarity. As observed, pianists within the same stylistic category exhibit consistently smaller DTW distances. For instance, the classical pianists (P1, P4, P7, and P10) form a cluster with relatively low pairwise distances, indicating similar rhythmic characteristics. In contrast, larger distances are observed between pianists from different stylistic groups, reflecting substantial differences in rhythmic expression.

Figure 7(b) illustrates the results of multidimensional scaling (MDS), which projects the high-dimensional DTW distance matrix into a two-dimensional space for visualization. The resulting distribution reveals clear clustering patterns corresponding to different performance styles. Specifically, the classical, romantic, and modern groups are well separated, demonstrating distinct rhythmic signatures across styles. These findings confirm that the proposed representation effectively captures stylistic differences in rhythmic evolution.

E. TEMPORAL STRUCTURE CONSISTENCY

The Temporal Structure Consistency Index assesses whether the rhythmic paths generated by the model exhibit stable patterns across different time scales. Specifically, this index calculates the autocorrelation coefficient of path segments to quantify the temporal repeatability and periodicity of the rhythmic state vector. A higher autocorrelation coefficient indicates stronger internal structure and consistency in rhythmic evolution, demonstrating that the model can

effectively capture recurring temporal patterns in the performance.

This index quantitatively evaluates the model's ability to learn and represent the inherent temporal organization of rhythms, while ensuring that the generated paths maintain smooth and expressive consistency. Specific evaluation data are shown in Table 1.

TABLE 1. Temporal Structure Consistency Indicators of Rhythmic Paths

Composer/Model	Intro	Theme A	Bridge	Theme B	Development	Recap	Coda	Average Autocorrelation
Bach	0.92	0.85	0.78	0.88	0.75	0.90	0.94	0.86
Mozart	0.76	0.68	0.72	0.65	0.70	0.74	0.79	0.72
Beethoven	0.81	0.75	0.80	0.77	0.73	0.82	0.86	0.79
Chopin	0.68	0.72	0.65	0.70	0.67	0.71	0.75	0.70
Debussy	0.88	0.82	0.85	0.80	0.83	0.87	0.91	0.85

Bach's Intro section exhibits high autocorrelation (0.92), reflecting stable rhythmic structure. Chopin's lower average (0.70) suggests freer, more flexible rhythm. Mozart and Beethoven display intermediate values (0.72 and 0.79), while Debussy's overall average (0.85) underscores coherence and artistic expressiveness.

F. RHYTHMIC DYNAMIC FEATURE FIDELITY

This metric evaluates the extent to which the model output captures dynamic rhythmic characteristics, including tempo variations, accelerations and decelerations, and expressive timing deviations such as rubato. Specifically, the statistical distributions of these dynamic features in the model-generated rhythm paths are compared with those derived from annotated reference data using the Kullback–Leibler (KL) divergence. This measure quantifies the discrepancy between the two distributions at a global level.

Lower KL divergence values indicate higher fidelity, reflecting closer agreement between the modeled and reference distributions of rhythmic dynamics. High fidelity

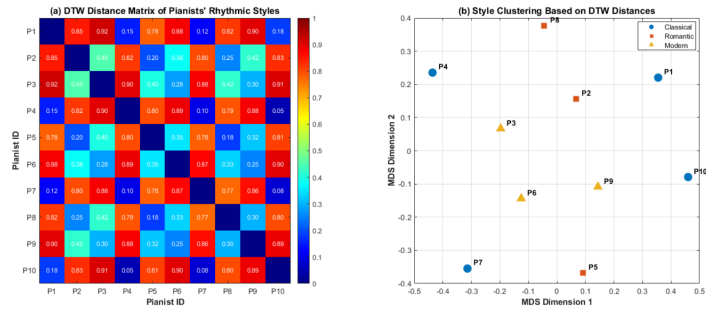


FIGURE 7. Analysis of Performance Style Differences: (a). DTW distance matrix of pianists' rhythmic styles; (b). MDS style clustering based on DTW distance

suggests that the model effectively preserves the expressive temporal characteristics present in the original performance.

This metric therefore provides evidence that the generated rhythmic paths not only resemble the reference trajectories in a structural sense but also faithfully reproduce the underlying expressive variations in rhythm. As shown in Table 2, the evaluation results demonstrate the fidelity of the model in capturing dynamic rhythmic features. The comparison includes both the mean values and the distributional differences between the model outputs and the manually annotated reference data, with the latter quantified using KL divergence. For tempo variation, the model achieves a mean value of 0.83, which is close to the reference mean of 0.85, with a KL divergence of 0.045, indicating strong agreement and effective preservation of tempo dynamics. Similarly, the KL divergence values for acceleration and dynamic range are 0.038 and 0.041, respectively, demonstrating high fidelity in modeling these temporal characteristics.

TABLE 2. Rhythmic Dynamic Feature Fidelity Evaluation

Dynamic Feature	Manual Annotation Mean	Model Output Mean	KL Divergence	Description
Tempo Change	0.85	0.83	0.045	Good fitting of tempo variations
Acceleration	0.60	0.58	0.038	Accurate capture of accelerations
Rubato	0.72	0.69	0.052	High preservation of expressivity
Expressivity				
Dynamic Range	0.77	0.75	0.041	Dynamic variations closely matched

The KL divergence for rubato expressiveness is slightly higher (0.052), suggesting comparatively greater variability in capturing expressive timing deviations. Nevertheless, the value remains low overall, indicating that the model retains a high level of accuracy in representing expressive rhythmic variations.

G. STABILITY AND ROBUSTNESS

Stability is defined as the degree of similarity between output rhythm paths under varying performance samples and noise conditions. To evaluate this, controlled perturbations are introduced into the input data, such as small random shifts in the onset timing of MIDI note events. The resulting output paths are then compared, and their average similarity is used as a measure of model stability.

A high level of robustness indicates that the model can reliably extract rhythmic evolution features without being overly sensitive to minor input variations. This metric therefore reflects the model's reliability in practical scenarios involving diverse performances and environmental noise. It further demonstrates the model's potential for generalization and its applicability to real-world rhythmic analysis tasks.

Table 3 summarizes the evaluation results of rhythmic path stability and robustness under different test conditions. Under baseline conditions, the model achieves a similarity score of 0.92 with a standard deviation of 0.03, indicating high stability and consistency.

TABLE 3. Stability and Robustness of the Rhythm Path

Test Condition	Average Path Similarity	Standard Deviation	Robustness Rating	Notes
No Perturbation (Baseline)	0.92	0.03	High	Stable model output
MIDI Timing Shift ± 5 ms	0.88	0.04	High	Insensitive to minor timing shifts
MIDI Timing Shift ± 10 ms	0.83	0.06	Moderate	Slight decline but overall stable
Different Performance Samples	0.85	0.05	High	Adapts well to varied performances
Environmental Noise	0.79	0.07	Moderate-High	Some tolerance to noise interference

When MIDI onset timing is perturbed by ± 5 ms, the similarity slightly decreases to 0.88, while robustness remains high, demonstrating the model's insensitivity to minor temporal variations. With increased perturbation of ± 10 ms, the similarity further declines to 0.83, accompanied by increased variability, indicating moderate degradation in performance while maintaining overall stability.

For variations across different performance samples, the model achieves a similarity score of 0.85, suggesting good adaptability to diverse playing styles. Under ambient noise conditions, the similarity decreases to 0.79 with a standard deviation of 0.07, indicating a certain degree of sensitivity to noise but still maintaining moderate-to-high robustness.

Overall, these results demonstrate that the proposed model exhibits strong stability under minor perturbations and maintains reasonable robustness under more challenging conditions, confirming its suitability for real-world rhythmic analysis scenarios.

V. CONCLUSION

This study presents a systematic approach for modeling the dynamic evolution of rhythmic characteristics in piano performance, where rhythm adapts fluidly to performance context. The Conv-TCN framework enables accurate extraction of temporal rhythm features from piano MIDI data, with potential application to modeling material stresses and flow in high-speed processes for defect prediction and prevention. Local short-term rhythmic patterns are extracted via 1D-CNN, while dilated TCN captures temporal and long-term dependencies. Multi-scale fusion integrates local and global information, producing expressive rhythm path sequences. Temporal normalization and smoothing enhance continuity and interpretability. The Conv-TCN model achieves superior accuracy (MSE = 0.032) and path smoothness (0.020) compared to LSTM and Transformer models. This research surpasses static rhythm analysis, representing the first systematic attempt to model predictive paths of rhythmic evolution in piano performance. The contributions extend to musicology and suggest adaptability for real-time diagnostic analysis and control in industrial applications, such as manufacturing. Future research may expand the model to multi-instrument concertos and realtime rhythm analysis, facilitating complex, synchronized control frameworks across multiple machines in continuous processing chains.

AUTHOR CONTRIBUTIONS

Conceptualization – Xiaolan Qiu; methodology – Xiaolan Qiu, Buwen Hou and Xin Wang; investigation – Xiaolan Qiu and Xin Wang; writing-original draft preparation – Xiaolan Qiu, Buwen Hou and Xin Wang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] W. G. Chen, J. R. Iversen, M. H. Kao, P. Loui, A. D. Patel, R. J. Zatorre, et al., "Music and brain circuitry: Strategies for strengthening evidence-based research for music-based interventions," *Journal of Neuroscience*, vol. 42, no. 45, pp. 8498-8507, 2022, doi: 10.1523/JNEUROSCI.1135-22.2022.
- [2] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1-39, 2023, doi: 10.1145/3597493.
- [3] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, et al., "Schubert Winterreise dataset: A multimodal scenario for music analysis," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 2, pp. 1-18, 2021, doi: 10.1145/3429743.
- [4] R. Su, "Piano teaching-assisted beat recognition based on spatio-temporal two-branch attention," *International Journal of Information and Communication Technology*, vol. 26, no. 5, pp. 100-116, 2025, doi: 10.1504/IJICT.2025.145149.
- [5] Y. Su and Y. Wang, "Optimization of music education strategy guided by the temporal-difference reinforcement learning algorithm," *Soft Computing*, vol. 28, no. 13, pp. 8279-8291, 2024, doi: 10.1007/s00500-024-09631-0.
- [6] Z. Liu, "AI-Driven classification and trend analysis of piano music genres using large language models," *International Journal of Information and Communication Technology*, vol. 26, no. 12, pp. 32-48, 2025, doi: 10.1504/IJICT.2025.146164.
- [7] Y. Jiang and Z. Sun, "Multimodal Sentiment Perception for Intelligent Music Generation Using Improved Transformer Architectures," *Informatica*, vol. 49, no. 5, pp. 147-166, 2025, doi: 10.31449/inf.v49i5.6864.
- [8] L. Wang, Z. Zhao, H. Liu, J. Pang, Y. Qin, and Q. Wu, "A review of intelligent music generation systems," *Neural Computing and Applications*, vol. 36, no. 12, pp. 6381-6401, 2024, doi: 10.1007/s00521-024-09418-2.
- [9] Z. Xiao, X. Chen, and L. Zhou, "Music performance style transfer for learning expressive musical performance," *Signal, Image and Video Processing*, vol. 18, no. 1, pp. 889-898, 2024, doi: 10.1007/s11760-023-02788-5.
- [10] D. V. T. Le, L. Bigo, D. Herremans, and M. Keller, "Natural language processing methods for symbolic music generation and information retrieval: A survey," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1-40, 2025, doi: 10.1145/3714457.
- [11] J. Min, Z. Gao, L. Wang, and A. Zhang, "Application research of short-time Fourier transform in music generation based on the parallel WaveGan system," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 9, pp. 10770-10778, 2024, doi: 10.1109/TII.2024.3397344.
- [12] P. Georges and A. Seckin, "Music information visualization and classical composers discovery: An application of network graphs, multidimensional scaling, and support vector machines," *Scientometrics*, vol. 127, no. 5, pp. 2277-2311, 2022, doi: 10.1007/s11192-022-04331-8.
- [13] F. Liu, D. L. Chen, R. Z. Zhou, S. Yang, and F. Xu, "Self-Supervised music motion synchronization learning for music-driven conducting motion generation," *Journal of Computer Science and Technology*, vol. 37, no. 3, pp. 539-558, 2022, doi: 10.1007/s11390-022-2030-z.
- [14] H. Jia, "Piano performance techniques and musical expressiveness," *Pacific International Journal*, vol. 6, no. 4, pp. 34-37, 2023, doi: 10.55014/pij.v6i4.461.
- [15] L. Peng, "Piano players' intonation and training using deep learning and mobilenet architecture," *Mobile Networks and Applications*, vol. 28, no. 6, pp. 2182-2190, 2023, doi: 10.1007/s11036-023-02175-x.
- [16] Y. Ghatas, M. Fayek, and M. Hadhoud, "A hybrid deep learning approach for musical difficulty estimation of piano symbolic music," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 10183-10196, 2022, doi: 10.1016/j.aej.2022.03.060.
- [17] A. Danielsen, R. Brøvig, K. K. Bøhler, G. S. Câmara, M. R. Haugen, E. Jacobsen, et al., "There's more to timing than time: Investigating musical microrhythm across disciplines and cultures," *Music Perception: An Interdisciplinary Journal*, vol. 41, no. 3, pp. 176-198, 2024, doi: 10.1525/mp.2024.41.3.176.
- [18] Y. Li and R. Sun, "Innovations of music and aesthetic education courses using intelligent technologies," *Education and Information Technologies*, vol. 28, no. 10, pp. 13665-13688, 2023, doi: 10.1007/s10639-023-11624-9.
- [19] K. Tsioutas and G. Xylomenos, "On the impact of audio characteristics to the quality of musicians' experience in network music performance," *J Audio Eng Soc*, vol. 69, no. 12, pp. 914-923, 2021.
- [20] A. Stevenson, "The performance of machine aesthetics: acoustic reimaging of electronic music," *Popular Music*, vol. 42, no. 4, pp. 438-456, 2023, doi: 10.1017/S026114302400014X.
- [21] X. Song and P. Phokha, "Applications of Artificial Intelligence-Assisted Computing In Piano Education," *Journal of Ecohumanism*, vol. 3, no. 7, pp. 1648-1659, 2024.
- [22] M. Clayton, S. Tarsitani, R. Jankowsky, L. Jure, L. Leante, R. Polak, et al., "The interpersonal entrainment in music performance data collection," *Empirical Musicology Review*, vol. 16, no. 1, pp. 65-84, 2021, doi: 10.18061/emr.v16i1.7555.