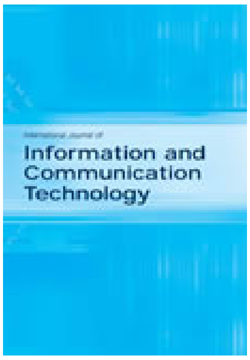
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



Information and
Communication
Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Piano performance beat assessment: integrating transformer with multimodal feature learning

Jun Deng

DOI: 10.1504/IJICT.2025.10074514

Article History:

Received:	12 August 2025
Last revised:	26 September 2025
Accepted:	26 September 2025
Published online:	20 November 2025

Piano performance beat assessment: integrating transformer with multimodal feature learning

Jun Deng

School of Arts,
Shandong Management University,
Jinan, 250000, China
Email: 14438120220412@sdmu.edu.cn

Abstract: This paper proposes PianoTrans-Fusion, a piano performance beat assessment system that integrates the transformer architecture with multimodal feature learning. The system uses three modalities, including audio, video, and MIDI, to perform feature extraction and preprocessing, respectively, and captures fine-grained temporal dependencies in the performance rhythm through multimodal fusion strategies and transformer-based processing modules. Comparative experiments on the MAESTRO dataset show that PianoTrans-Fusion improves rhythm consistency to 0.032 and reduces beat error to 0.071 compared to five baseline methods. Ablation experiments further verify the key roles of transformer, multimodal fusion, and self-attention mechanisms. The results indicate that the system has advantages in terms of accuracy and robustness in beat evaluation, and has application value in intelligent piano accompaniment, music education, and automated performance feedback.

Keywords: transformer; multimodal feature learning; piano performance; beat assessment.

Reference to this paper should be made as follows: Deng, J. (2025) ‘Piano performance beat assessment: integrating transformer with multimodal feature learning’, *Int. J. Information and Communication Technology*, Vol. 26, No. 41, pp.74–90.

Biographical notes: Jun Deng obtained her Doctoral degree from Anyang University in South Korea in 2022. She is currently a Lecturer at the School of Arts, Shandong Management University. Her research interests include music education, music therapy and multimodal feature learning.

1 Introduction

In recent years, the development of artificial intelligence (AI) and deep learning (DL) has promoted the application of intelligent music analysis, which has attracted increasing attention, especially in music teaching, composition, and automatic performance (Han, 2025). Piano performance is not only a display of technique, but also rich artistic expression, and beat is the foundation of this, only when the rhythm is steady can the performance be smooth and expressive. But most traditional ways of measuring beats depend on manual observation or old signal processing technologies. These approaches are not only slow and subjective, but they also don't match the needs of real-time and intelligence.

Many systems try to automatically assess rhythm with AI. Traditional audio analysis uses pitch and duration to establish rhythm, but it is susceptible to noise and cannot capture small performance changes. Later models like convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs) have improved this, but most focus solely on audio, ignoring visual and musical instrument digital interface (MIDI) information like hand movements and key presses during performance, limiting assessment accuracy (Qiu et al., 2021).

To address this, this paper proposes a multimodal beat assessment method that combines audio, video, and MIDI data, based on the transformer model, allowing different signals to complement each other. The system automatically determines which type of information is more reliable through an attention mechanism and dynamically fuses them to capture the rhythm of the performance more precisely. The goal is to provide a more realistic and sensitive technical solution for intelligent practice and teaching feedback.

We want to make the automation and intelligence of piano performance beat assessment better through this research. We also want to find new ways to combine multimodal features, transformer and give more accurate and reliable technical support for things like music education, intelligent practice, and automated performance. This field is predicted to have more applications for personalised music education and performance feedback systems in the future, thanks to the addition of more modal data and advances in technology.

2 Relevant work

2.1 Piano performance beat assessment method

Piano performance beat assessment is an important research direction in the field of music information processing, especially in applications such as intelligent piano education, automated performance feedback systems, and intelligent accompaniment, where it has broad potential. Consequently, the effective evaluation of beat in piano playing has emerged as a prominent research focus in both academic and industrial spheres in recent years. There are two main types of methods for evaluating the beat of piano performances: signal processing-based methods and machine learning (ML)-based methods. As technology keeps becoming better, the latter has become the norm.

In the beginning, beat assessment approaches that used signal processing were the most popular. Most traditional ways of processing signals use the time domain or frequency domain properties of audio signals to figure out where the beat is (Melo et al., 2020). The amplitude envelope analysis method, for instance, figures out where the beat is by looking at how the amplitude of the audio input fluctuates. Although this method is relatively simple, it often fails to maintain good robustness and accuracy when faced with complex musical works and environmental noise. To overcome these problems, researchers have also tried using the autocorrelation function method, which detects rhythm cycles and infers beat positions by calculating the correlation between audio signals and their delayed versions.

Another conventional method is the beat detection method based on frequency domain analysis. This method commonly uses Fourier transformation to change the audio input into frequency domain information so that the rhythm features may be studied. This approach works well for separating distinct frequency components in an audio source and is good for picking up low-frequency rhythm information. But frequency domain analysis doesn't work very well when there are complicated chords, quick notes, or auditory interference (de Cheveigné, 2021). While these methods laid the groundwork for initial beat assessment studies, they frequently encounter difficulties in managing intricate rhythm patterns and transitions, particularly in rapid, multi-note performances and multi-track contexts, when precision is markedly diminished.

With the advent of ML technology, the evaluation of piano performance has progressively transitioned from conventional signal processing techniques to data-driven ML methodologies. ML approaches learn from a lot of performance data and can better pick out features from multimodal data including audio signals, MIDI data, and video signals to make predictions and evaluations of beats. Support vector machine (SVM) and decision tree (DT) are two traditional ML approaches that have been widely used for beat position categorisation and rhythm pattern detection (Subba and Chingtham, 2024). SVM finds the best hyperplane for classification by mapping input characteristics to a high-dimensional space. DT, on the other hand, builds a hierarchical framework to group distinct rhythm patterns. While these methods may yield efficient solutions in straightforward settings, they generally struggle with intricate rhythm patterns, particularly in the context of prolonged sequences and diverse, evolving playing styles, where their efficacy is markedly constrained.

As DL has become more popular, beat assessment algorithms that use CNNs have slowly started to appear. CNNs can get high-level time-frequency features from raw audio data, and they work well for music information processing jobs because they have been used successfully for image recognition (Gupta et al., 2022). When it comes to beat assessment, CNNs can automatically pull out useful rhythm characteristics from the audio spectrogram without needing features that were made by hand. One way to use CNNs to find beats is to turn the audio input into a Mel spectrogram and use convolution and pooling layers to get features that can properly capture information like note duration, pitch, and intensity. But CNNs aren't strong at processing long-term dependencies and have trouble capturing global rhythmic patterns, thus they aren't good at handling rhythmic patterns that change all the time.

Convolutional recurrent neural network (CRNN) combines the local feature extraction capabilities of CNN with the temporal modelling capabilities of RNN, giving it a natural advantage when processing multimodal data. CRNN can get time-frequency information from audio signals and temporal features through the RNN module. This makes it very good at finding beats. Although CRNN has achieved significant results in beat assessment, in practical applications, how to further improve its real-time performance and robustness, especially when multimodal data is missing or noisy, remains an urgent issue to be addressed.

As DL architecture based on self-attention mechanisms, transformer models have gradually made significant progress in various time series tasks. The main benefit of the transformer is that it can effectively capture long-range temporal connections without running into the vanishing gradient problem that typical RNNs have when modelling long-term dependencies. The transformer also uses self-attention techniques to do parallel computing, which not only makes the computations more efficient but also looks at the links between multiple input features at the same time (Wahid et al., 2023). The transformer model can better understand how different time steps in audio, MIDI, and video signals are related to each other when it comes to piano performance beat assessment. It does especially well at combining data from different sources and modelling long time sequences. As multimodal data processing technology advances, research that integrates diverse modal information for beat evaluation is poised to emerge as a significant trend in the future.

2.2 Transformer model

Vaswani et al. came up with the self-attention mechanism in 2017, which is the basis for the transformer model (Meel and Vishwakarma, 2023). Unlike traditional RNNs and CNNs, transformers rely entirely on self-attention mechanisms to capture temporal dependencies in sequences, without using traditional recursive or convolutional structures.

The transformer uses self-attention to dynamically calculate position-relationships when processing input data. By capturing information from multiple points across the sequence, this technique avoids gradient vanishing or exploding difficulties that typical RNNs may have when processing long-term sequences. The transformer also uses a multi-head attention mechanism to parallelly compute alternative attention representations to capture multiple information patterns in input data.

Encoder and decoder make up the transformer. Each encoder and decoder module has numerous identical sub-layers, mostly self-attention and feedforward neural network layers. The encoder extracts information from the input sequence, while the decoder creates the target sequence from its output. Each self-attention layer generates attention weights for each sequence position, determining the degree of linkage between elements (Wei et al., 2021). Layering improves the model's grasp and abstraction of input data. The model uses query, key, and value to calculate the self-attention mechanism's associations between input sequence elements. Fundamental formula of self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V represent the query, key, and value sets, respectively, and d_k is the dimension of the key. With this formula, the transformer figures out an attention weight set based on how the query and keys are related. It then uses that set on the value set to get the weighted output.

In practical applications, the transformer further enhances its ability to understand information through a multi-head attention mechanism. The multi-head attention mechanism maps queries, keys, and values to multiple subspaces, independently calculates attention in each subspace, and finally concatenates the outputs of multiple attention heads to form a unified representation (Ren et al., 2022). This mechanism enables the transformer to simultaneously focus on different aspects of the input sequence, thereby better understanding and modelling the complex structure of the input data.

2.3 Multimodal feature learning

Information in the real world is often diverse: it includes images, sounds, text, and sensor signals. A single source can be biased but combining them allows the model to see more clearly and make more stable judgements. In video analysis, for instance, the visual modality gives spatial information about the scene, the audio modality gives temporal information about sounds and voice, and the text modality adds semantic-level information. This is the core idea behind multimodal feature learning, instead of looking at just one type of data, the model learns to integrate different sensory inputs.

Multimodal feature learning relies on fusing input from different modalities. Modality differences hinder feature fusion because text is a discrete sequence of symbols, images are high-dimensional pixel sets, and audio is a time-series signal. Researchers have developed many mainstream fusion solutions to address this issue. Feature-level fusion directly concatenates, or transfers feature variables from distinct modalities into a shared subspace for joint modelling (Zhao et al., 2024). This method is straightforward to apply and integrates information, although feature scale discrepancies between modalities may weaken or lose certain modal information. Thus, performance improvement generally requires modal feature preprocessing or weighted fusion. Decision-level fusion also uses weighted voting or probabilistic fusion to combine the outputs of each modality after training models for them independently. This technique optimises by selecting the best algorithm for each modality, but it often overlooks deep interactions and linkages between modalities, potentially missing latent connections.

Multi-modal learning approaches using deep neural networks (DNNs) have become common as DL technology advances. By concurrently training deep networks, these approaches may automatically extract and integrate data from diverse modalities. Two-stream networks improve task performance by designing distinct neural network branches for each modality and fusing features at higher levels (Xiong et al., 2020). Attention processes allow models like transformers to dynamically assign weights to modalities, making fusion more flexible. Cross-modal generative networks are also used more. In lacking or noisy environments, these networks fuse features across modalities and produce data for missing modalities, improving system robustness.

The main benefits of multimodal feature learning are information complementarity, model resilience, and semantic richness. Models can transfer information across modalities to improve the system's task understanding by combining features. Multimodal sentiment analysis combines facial expressions in photos, voice in audio, and lexical information in text to better assess an individual's mood. Despite noise or loss of features from one modality, other modalities can still contribute enough information to the model, boosting system robustness.

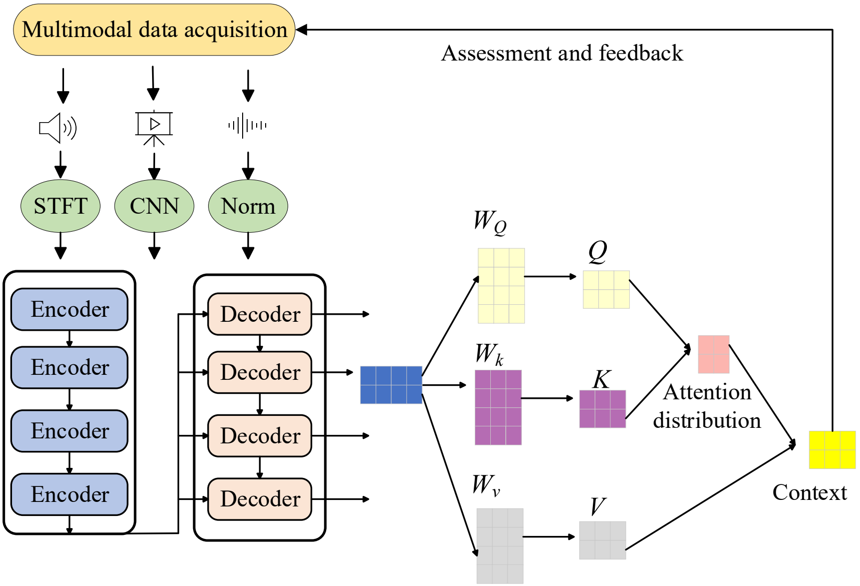
As multimodal learning technology advances, forthcoming study will predominantly concentrate on many domains. On one hand, model lightweighting and efficiency are important, especially when resources are limited. Techniques like model compression and distillation can help with this by lowering the computational load and speeding up real-time performance. On the other hand, self-supervised and weakly supervised learning use unlabelled or weakly labelled data for cross-modal pre-training to reduce the need for labelled data (Ericsson et al., 2022). Also, research on cross-modal generation and inference is slowly becoming available. This research uses generative models to fill in the gaps in missing modalities, which makes the model more stable when multimodal input is incomplete or noisy. As technology gets better, cross-domain generalisation will also get stronger, which will help multimodal learning work better in a wider range of tasks and settings.

In short, multimodal feature learning can make model work better and be more stable by combining input from several sources. This is especially true for complicated tasks and contexts, where it greatly improves the model's grasp of the target. As DL keeps becoming better, especially with new ideas in attention mechanisms and generative modelling, we'll look into further ways that multimodal learning might be used in many industries.

3 System design and methods

The PianoTrans-Fusion piano performance beat assessment system suggested in this study is built on the transformer model and learning from many different types of features. Figure 1 shows that the goal is to use audio, video, and MIDI data together to help the performer improve by giving them correct beat assessments and real-time feedback.

Figure 1 System framework for piano performance beat assessment (see online version for colours)



3.1 Data acquisition module

The piano performance beat assessment system is built on the data acquisition module. It is in charge of making sure that data from diverse source is collected at the same time and that the data from different sources is always in the same time zone. The system uses a single set of technical tools to gather data from three different types of input: audio, video, and MIDI. This lets it fully and correctly records the rhythm and finger movements of piano playing. High-sampling-rate microphones collect audio data, standard cameras record the performer's movements, and MIDI data immediately gives us digitised performance data.

To get enough frequency resolution to pick up the subtleties of piano notes, audio signals are digitised at a sampling frequency of 44.1 kHz. The system makes a spectrogram by using short-time Fourier transform (STFT) to look at the audio signal's frequency and rhythm (Min et al., 2024). The following formula can be used to describe how the spectrogram is made:

$$S(t, f) = \sum_{n=\infty}^{\infty} x(n) w(n-t) e^{-j2\pi fn} \tag{2}$$

where $S(t, f)$ stands for the spectrogram, $x(n)$ stands for the discrete representation of the audio signal, $w(n - t)$ is the window function, f is the frequency, and t is the time point. The STFT gives the audio stream a time-frequency representation that shows the spectral characteristics of piano notes. This helps find rhythmic patterns.

A regular camera records the performer's finger motions in video data. The system employs CNN to find hands and trace their paths in real time so it can get information about finger movement on the keyboard. Video data is mostly utilised to record the performer's finger placements and activity sequences and look at how they relate to rhythm (Clayton et al., 2020). This step is very important for real-time feedback since it lets the system see variations in pace and exact finger motions during the performance.

MIDI signals show how a piano is played in digital form. Each MIDI event has notes, pitch, velocity, and timestamps. The system reads the MIDI data to find out when each note starts and ends. These times are used to figure out the rhythm. The system uses the following formula to standardise the timestamps of MIDI events:

$$T_m = \frac{t}{f_m} \tag{3}$$

where T_m is the time stamp for the MIDI event, t is the actual time the event happens, and f_m is the MIDI signal's sampling frequency. To make sure the tempo data is correct, the machine samples it 1,000 times per second.

3.2 Feature extraction and preprocessing module

The feature extraction and preprocessing module is responsible for extracting key features from audio, video, and MIDI data, and performing noise reduction, standardisation, and normalisation to improve data quality. Its core objective is to unify multimodal information into a single feature space for subsequent fusion.

For audio signals, the system first converts them into a time-frequency representation using STFT to capture changes in note frequency. Based on this, Mel frequency cepstral coefficients (MFCC) are extracted to characterise timbre and rhythm features (Hawi et al., 2022). MFCC is sensitive to human auditory perception and is a commonly used metric in audio analysis. The calculation formula is as follows:

$$MFCC(n) = \sum_{m=1}^{M} \log \left| X_m(n) \right| \cdot W_m(n) \tag{4}$$

where $MFCC(n)$ stands for the n^{th} Mel-frequency cepstral coefficient, $X_m(n)$ stands for the amplitude spectrum of the m^{th} frequency band in the spectrogram, and $W_m(n)$ stands for the filter bank weights for the Mel frequencies. The system can get more compact and rhythm-related audio features using MFCC, which makes it easier to judge the beat later on.

The system employs CNN to extract features from video data so it can figure out where the performer's fingers are and what they are doing. The CNN network processes each video frame into a feature vector that shows the locations of important spots on the hand (Fadl et al., 2021). The method uses backdrop removal and high-pass filtering to pre-process video frames. This makes finger movement features stand out more and reduces noise. The system can effectively extract temporal information related to piano playing rhythm by analysing video frames in a time series.

When it comes to MIDI data, the system gets input information including the timestamp, note, and velocity of each MIDI event. The system standardises the pitch, length, and velocity of each note during the preprocessing of MIDI data (Jeong et al., 2020). It then changes these properties into time-series data so that they may be used with audio and visual data. The system takes the timestamp of each MIDI event and the note information and combines them using the following formula:

$$T'_m = \frac{T_m - \min(T)}{\max(T) - \min(T)} \tag{5}$$

where T'_m is the normalised MIDI timestamp, T_m is the original timestamp, min(T) and max(T) are the lowest and highest values in the MIDI timestamp sequence, respectively. By normalising the temporal information of all MIDI events, the system makes sure that it can be processed on a single scale. This stops biases that can come from disparities in temporal spans.

During the preprocessing stage, all data is standardised and normalised so that information from different modalities can be compared and combined on the same scale. After preprocessing, audio, video, and MIDI features can all go into the multi-modal feature fusion module for more weighted feature fusion.

3.3 Multimodal feature fusion module

In rhythm assessment, relying solely on audio, video, or MIDI is insufficient. To address this, the system combines the features of all three signal types to complement one another. For example, audio may be affected by environmental noise, but video can provide clues through hand movements; MIDI signals are precise but occasionally missing and can be corrected using audio.

The fusion process is not a simple averaging but rather allows the model to determine when to rely more on audio and when to focus on video. Specifically, given the features of audio (A), video (V), and MIDI (M), the model first calculates the degree of attention between them using the following formula:

$$Q_i = W_Q \cdot x_i \tag{6}$$

$$K_i = W_K \cdot x_i \tag{7}$$

$$V_i = W_V \cdot x_i \tag{8}$$

where x_i represents the input feature of the i^{th} modality, W_Q, W_K and W_V are the weight sets of query, key, and value, respectively, Q_i, K_i and V_i are the query, key, and value variables transformed by these weight sets through the modality features.

Next, the system performs weighted fusion of the audio, video, and MIDI features through the calculated attention weights. The fused features are represented as:

$$F_{fused} = \sum_{i=1}^{3} \alpha_i \cdot F_i \tag{9}$$

where F_i is the feature of the i^{th} mode, α_i is the weight automatically assigned by the model, and F_{fused} is the final merged feature. Through weighted summation, each mode is dynamically combined according to its importance, allowing signals that are more helpful for judging the current beat to have a greater weight.

3.4 Transformer processing module

The purpose of this module is to use the transformer model to process the fused features even more in order to get useful rhythm information and, in the end, get rhythm evaluation findings.

The self-attention mechanism is the most important part of the transformer model. It does a great job of finding relationships between incoming data. In this module, the transformer uses several self-attention layers to process the multimodal fused characteristics. The model initially gets the Query, Key, and Value variables by doing a linear transformation on the input feature representation F_{fused}:

$$Q_i = W_Q \cdot F_{fused} \tag{10}$$

$$K_i = W_K \cdot F_{fused} \tag{11}$$

$$V_i = W_V \cdot F_{fused} \tag{12}$$

where F_{fused} shows the features after multimodal fusion, W_Q, W_K and W_V show the weight sets for the query, key, and value, in that order, Q_i, K_i and V_i are the new versions of the query, key, and value variables. The transformer can find the relationships between features from multiple modalities and give each input feature an adaptive weight by figuring out these factors.

Then, the transformer figures out the attention weights to get a weighted representation of each input feature. The system can figure out how important each feature is in the current context by figuring out how similar the query variable is to the key variable. It may then use this information to do a weighted summing to get the final representation of each modal feature.

The transformer processing module sends the output weighted features to the FFN after they have been processed through several layers of self-attention processes. The FFN uses a series of nonlinear modifications to get more rhythm information (Nieto-del-Amor et al., 2021). This helps the system grasp how the beat is structured in the performance. Residual connections and layer normalisation are used to improve the output features so that information can flow smoothly across the network and avoid problems with gradient vanishing.

The transformer is great because it can represent things on a global scale and understand dependencies in a flexible way. Transformer can thoroughly comprehend the temporal links between notes in a performance and pick up on the small variations in rhythm. This lets it efficiently and reliably analyses the beat of the piano play. Transformer can also manage complicated relationships between multiple modal aspects, which gives a more complete view of how to judge a beat.

3.5 Assessment and feedback module

The assessment and feedback module first looks at the output of the transformer model's features more closely, notably by looking at the rhythm's consistency, precision, and stability. To accurately assess the performer's beat performance throughout the performance, the system separates the evaluation process into two primary parts: rhythm consistency evaluation and rhythm accuracy evaluation.

The system's major focus for rhythm consistency assessment is the stability of the beat during the performance. This is how it decides if the artist has kept a steady rhythm. The system does this by calculating the standard deviation of the performance rhythm to find out how big the rhythm deviation is (Moon et al., 2023). To find the standard deviation σ, do the following:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - \mu)^2} \tag{13}$$

where T_i is the playing time of the i^{th} note, μ is the mean value of the note playing time, and N is the total number of notes. A smaller standard deviation means that the rhythm is more consistent, which means that the performer did a better job of keeping the rhythm.

To check how accurate the rhythm is, the system figures out the beat error, which is the difference between the performer's real performance and the ideal beat. To get the beat mistake, we compare the difference between the actual playing time of each note and the ideal beat time. The formula for doing this is as follows:

$$E_{beat} = \frac{1}{N}\sum_{i=1}^{N}\left|T_i - T_{ideal}\right| \tag{14}$$

where T_i is the real time for the i^{th} note, T_{ideal} is the best time for the beat, and N is the total number of notes. The performer has a better sense of rhythm if the beat mistake is reduced.

The system uses weighted summation to combine the scores for rhythm consistency and accuracy to get the final complete rhythm score, which is then given to the performer as feedback. This is how we can write the total rhythm score S_{total}:

$$S_{total} = \alpha \cdot S_{consistency} + \beta \cdot S_{accuracy} \tag{15}$$

where $S_{consistency}$ is the rhythm consistency score, $S_{accuracy}$ is the rhythm accuracy score, α and β are the coefficients that show the weights of consistency and accuracy, respectively. The system may easily change the balance between consistency and accuracy by changing α and β to fit the needs of different application scenarios.

Lastly, the evaluation findings are shown not just as numbers, but also as a graphical interface, audio feedback, or text feedback. This helps the performer find and fix rhythm faults while they practice. For instance, the system can show the artist a chart showing the beat error for each note they play, or it can provide them with audio alerts in real time to let them know when they are off the ideal beat.

4 Experimental results and analyses

4.1 Experiment data and settings

This study employed the MAESTRO dataset as the experimental dataset to validate the proposed PianoTrans-Fusion piano performance beat evaluation system. The MAESTRO dataset has a lot of audio, MIDI, and performer information on piano performances. It is good for jobs that involve combining several types of data and judging rhythm. The dataset contains high-quality audio and MIDI data that is correctly annotated, which makes sure that the assessment is accurate and reliable.

We chose a section of the MAESTRO dataset to use for training and testing in this experiment. The dataset contains performance data from several piano players, each of whom played a different piece of music with a varied rhythm and style of playing. For multimodal feature extraction, the audio and MIDI data are processed at the same time so that the data stays the same. Each audio file of the performance is lined up with the MIDI file that goes with it.

Table 1 shows the most important information of the MAESTRO dataset.

Table 1 Information on the MAESTRO dataset

Data item	*Description*
Number of audio files	1,000 piano pieces
Audio duration	Average duration of each piece is 3–4 minutes
Number of MIDI files	1,000 corresponding MIDI files with detailed performance data
Number of performers	16 pianists with varying skill levels
Data annotations	Each MIDI file contains detailed note timestamps, pitch, velocity, and other performance information
Use case	Audio analysis, rhythm evaluation, performance style analysis, etc.

This experiment mainly employs audio (in WAV format) and MIDI data, together with video data, to do a multimodal analysis to see if the PianoTrans-Fusion system can accurately measure rhythm across different types of data. We used the MAESTRO dataset to test how well the PianoTrans-Fusion system worked in this experiment. The data files were carefully lined up so that data fusion would work well when extracting features from many sources. The dataset was split into three parts: a training set (70%), a validation set (15%), and a test set (15%) which was done so that the model could be trained and tested (Śmigiel et al., 2021). This split makes sure that the model has enough training data to learn from and that it can also be tested and validated to see how well it generalises.

4.2 Comparative experiments

To validate the efficacy of the PianoTrans-Fusion system, this study devised a comparative experiment to evaluate the proposed system against current piano performance beat assessment methodologies. The comparative experiment showed how good the PianoTrans-Fusion system is by comparing the performance of different approaches based on how well they kept the rhythm and how accurate they were. To fully assess the system's performance, we chose the following five baseline approaches for comparison:

- Traditional audio-based beat estimation methods: this method uses time-domain or frequency-domain analysis to figure out the beat by taking features from audio signals. This method was commonly employed in initial rhythm evaluation; however, it primarily focuses on audio signals and fails to capture visual and MIDI data from the performer.
- LSTM-based audio rhythm estimation method: this approach employs LSTM to guess beats based on sequences of audio features. LSTM can work with time-based data, but it only uses audio signals and can't make full use of information from other types of data.
- Single-modal transformer model: this technique relies on the transformer model and exclusively utilises audio signals as input. Even though transformer can accurately capture long-term dependencies, this method doesn't include other types of data, which could lead to inaccurate rhythm evaluation.
- Audio-MIDI based method: this approach employs both audio and MIDI data to check the rhythm. It combines MIDI note timestamps and pitch information with audio elements to make the rhythm check more accurate. The use of multi-modal information improves the accuracy of assessments; however, this method doesn't include video data, which limits its effectiveness.

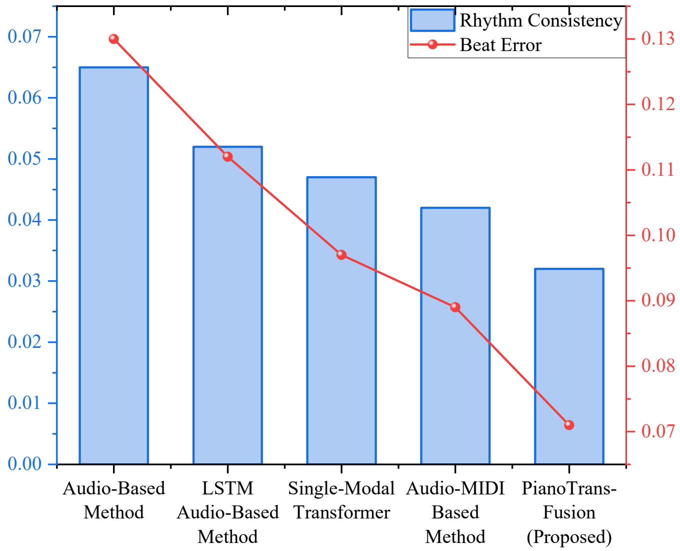
We tested the five approaches on the same training, validation, and test sets, especially looking at two things: how consistent the rhythm was and how many beat errors there were (Torres-Soto and Ashley, 2020). To get the standard deviation between notes, we can measure rhythm consistency. To find the difference between the actual performance time and the ideal beat time, we can measure beat error. Figure 2 shows the outcomes of the experiment.

The comparative experiment's results demonstrate that the PianoTrans-Fusion system is better than all other approaches in two important areas: rhythm consistency and beat inaccuracy.

The PianoTrans-Fusion system demonstrates outstanding performance in terms of rhythm stability, significantly outperforming other methods. Compared to traditional Audio-Based Method, it reduces rhythm fluctuations by approximately half, indicating that it restores performance rhythms in a more stable and natural manner.

In terms of beat accuracy, PianoTrans-Fusion demonstrated stronger control capabilities, with a measured error of only 0.071. In contrast, the error of the traditional audio-based method reached 0.130, while the improved LSTM audio-based method only reduced it to 0.112, which is still significantly high. Although single-modal transformer or audio-based method have also made progress in rhythm restoration, such as reduced fluctuations and more consistent beats, there is still a significant gap in overall accuracy between them and PianoTrans-Fusion.

Figure 2 Results of comparative experiments (see online version for colours)



The conventional audio-based method and LSTM audio-based method were not very good compared to other approaches. This is mostly because they only use audio features and can't completely use other types of data to fix and add to faults that could happen in the audio. The single-modal transformer brings the benefits of the transformer model, which is good at finding long-term relationships in audio data. However, it still doesn't work as well as multi-modal systems because it only uses audio data. The audio-MIDI based method makes rhythm evaluation more accurate by merging audio and MIDI data. However, it still doesn't take into account visual information, which makes it less reliable for rhythm consistency and beat precision.

In conclusion, the PianoTrans-Fusion system surpasses current comparison approaches in rhythm consistency and accuracy by effectively integrating multi-modal feature fusion with the transformer model. This demonstrates the significance of multi-modal data fusion and self-attention mechanisms in evaluating piano performance beats, while also validating the efficacy and benefits of the PianoTrans-Fusion system introduced in this study for practical applications.

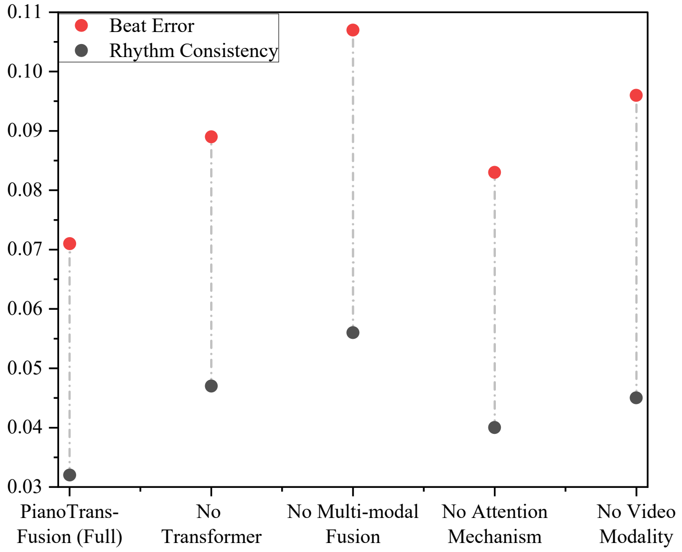
4.3 Ablation experiments

We did ablation tests to better understand how each part of the PianoTrans-Fusion system works and how different modules affect rhythm assessment performance. We could easily see how each module affected the overall performance by slowly taking out or replacing some essential modules in the system. The ablation studies aimed to validate the impacts of the transformer model, multimodal feature fusion, and self-attention mechanism on rhythm assessment efficacy.

- No transformer: for this experiment, we took out the transformer processing module from the PianoTrans-Fusion system and just used classic feature extraction methods to process the audio, MIDI, and video fusion information. The goal of this experiment was to see if the transformer model makes rhythm evaluation work better in a big way. We substituted transformer with alternative ML models and monitored the variations in rhythm consistency and beat inaccuracy subsequent to the removal of this module.
- No multi-modal fusion: this means not using multi-modal feature fusion. We process audio, video, and MIDI data separately in this experiment, without combining features from other modes. We specifically leverage single-modal elements from audio, video, and MIDI to check the rhythm. This experiment enables the assessment of the enhancement provided by multi-modal feature fusion in rhythm evaluation, as well as the constraints of single-modal features in this context.
- No attention mechanism: for this experiment, we took off the self-attention mechanism from the transformer and utilised a simple weighted average method to combine features from different modalities instead. The goal of this experiment is to find out how important the self-attention mechanism is for processing information from more than one source. The self-attention system changes the weighting ratios in real time based on how important different modalities are. Taking this mechanism out could make the system work worse.
- No video modality: this experiment got rid of video data and only used audio and MIDI modalities to check the beat. This experiment looked at how the video modality helped with rhythm evaluation and whether the system's ability to assess rhythm would drop considerably without visual information.

We did a thorough examination of the effect of each module using the four experimental settings mentioned. Figure 3 shows the outcomes of the experiment.

Figure 3 Results of ablation experiments (see online version for colours)



The ablation experiments clearly illustrate that taking off any module makes the system work worse, which shows how important each module is to the PianoTrans-Fusion system.

Without the transformer module, the system's rhythm consistency and beat error both go down, but the rhythm consistency goes down more (from 0.032 to 0.047). This illustrates that transformer is vital for capturing long-term dependencies and dealing with complicated rhythm shifts. The system's capacity to simulate extended sequences goes down a lot if this module is taken out, which makes the evaluation less accurate.

When multimodal feature fusion was taken away, the system's rhythm consistency and beat error got a lot worse, notably beat error, which went from 0.071 to 0.107. This indicates that multimodal fusion is a crucial element in enhancing the precision of rhythm assessment. Single-modal data is inadequate for capturing nuanced rhythmic variations in performance, particularly in intricate performance contexts, where multimodal information might synergise to enhance the comprehensiveness and precision of the evaluation.

When the self-attention mechanism was taken out, the system's performance got worse, especially when it came to keeping the rhythm consistent (from 0.032 to 0.040). This demonstrates that during the processing of multimodal characteristics, the self-attention mechanism can dynamically modify the weights of various modalities to enhance the outcomes of rhythm evaluation. Without the self-attention mechanism, the system can't completely weigh each modality based on how important it is, which makes the assessment less accurate.

Taking away the video option, the system's rhythm assessment performance stayed good when the video modality was taken away. However, both rhythm consistency and beat error went up compared to the system that had the video modality. This indicates that the video modality plays a complementary role in capturing the performer's hand movements and visual features, particularly when assessing rhythm. Visual information helps the system better understand performance details and reduce errors in audio and MIDI data.

In summary, the ablation experiment results indicate that each module of the PianoTrans-Fusion system plays an important role in overall performance, particularly the transformer model, multimodal feature fusion, and self-attention mechanism. The collaborative work of these modules enables the system to accurately capture rhythm changes in piano performance, enhancing the accuracy and robustness of rhythm assessment.

5 Conclusions

This study proposes a system called PianoTrans-Fusion that integrates transformer and multimodal feature learning. Through multimodal feature fusion mechanisms and the self-attention structure of the transformer model, the system effectively captures rhythm changes and detailed features during the performance process, achieving consistency and accuracy in the evaluation of piano performance rhythm. This paper constructs five core modules of the system and conducts experimental verification based on the MAESTRO dataset. In comparative experiments, PianoTrans-Fusion outperforms other baseline methods in terms of rhythm consistency and beat error, proving the effectiveness of multimodal feature fusion and the transformer structure. In ablation studies, the removal of certain modules resulted in diminished performance, thereby substantiating the essentiality and contribution of each system component. In general, this study has made great strides in making piano performance beat assessment more accurate and reliable. It has also opened up new technical possibilities for intelligent piano accompaniment, music education, and automated performance feedback systems.

This study has yielded specific results; yet it remains subject to some constraints. The MAESTRO dataset is the main source of experimental data. This dataset is big and good, but the performance scenarios are quite boring. They don't include a lot of background noise or interference from other performance conditions, which could make it harder for the system to generalise to more complicated real-world situations. Second, the system still has a lot of processing power needed to handle multimodal data in real time, especially when the transformer structure works with long sequence data, which needs a lot of hardware power. This work also does not look closely at how errors in aligning different modalities over time affect beat evaluation outcomes, which could be a significant thing to think about in real-world situations.

Subsequent research may be pursued in the following avenues. To improve the system's ability to generalise and be strong, we can add more diverse datasets that cover different performance levels and situations. Second, we can look for lightweight transformer structures and effective multimodal feature fusion algorithms to speed up real-time rhythm assessment, which will make it easier to use the system on embedded devices. Finally, beat assessment could be combined with more advanced music understanding tasks like performance style analysis and emotion recognition. This would allow for a shift from basic rhythm assessment to a full analysis of music performance, and it would push the development of smart music analysis systems to new heights.

Declarations

All authors declare that they have no conflicts of interest.

References

Clayton, M., Jakubowski, K., Eerola, T., Keller, P.E., Camurri, A., Volpe, G. and Alborno, P. (2020) 'Interpersonal entrainment in music performance: theory, method, and model', *Music Perception: An Interdisciplinary Journal*, Vol. 38, No. 2, pp.136–194.

de Cheveigné, A. (2021) 'Harmonic cancellation – a fundamental of auditory scene analysis', *Trends in Hearing*, Vol. 25, p.23312165211041422.

Ericsson, L., Gouk, H., Loy, C.C. and Hospedales, T.M. (2022) 'Self-supervised representation learning: Introduction, advances, and challenges', *IEEE Signal Processing Magazine*, Vol. 39, No. 3, pp.42–62.

Fadl, S., Han, Q. and Li, Q. (2021) 'CNN spatiotemporal features and fusion for surveillance video forgery detection', *Signal Processing: Image Communication*, Vol. 90, p.116066.

Gupta, S.S., Hossain, S. and Kim, K-D. (2022) 'Recognize the surrounding: development and evaluation of convolutional deep networks using gammatone spectrograms and raw audio signals', *Expert Systems with Applications*, Vol. 200, p.116998.

Han, Y. (2025) 'Exploring a digital music teaching model integrated with recurrent neural networks under artificial intelligence', *Scientific Reports*, Vol. 15, No. 1, p.7495.

Hawi, S., Alhozami, J., AlQahtani, R., AlSafran, D., Alqarni, M. and El Sahmarany, L. (2022) 'Automatic Parkinson's disease detection based on the combination of long-term acoustic features and Mel frequency cepstral coefficients (MFCC)', *Biomedical Signal Processing and Control*, Vol. 78, p.104013.

Jeong, D., Kwon, T. and Nam, J. (2020) 'Note-intensity estimation of piano recordings using coarsely aligned midi score', *Journal of the Audio Engineering Society*, Vol. 68, Nos. 1/2, pp.34–47.

Meel, P. and Vishwakarma, D.K. (2023) 'Multi-modal fusion using Fine-tuned Self-attention and transfer learning for veracity analysis of web information', *Expert Systems with Applications*, Vol. 229, p.120537.

Melo, D.d.F.P., Fadigas, I.d.S. and Pereira, H.B.d.B. (2020) 'Graph-based feature extraction: a new proposal to study the classification of music signals outside the time-frequency domain', *PLoS One*, Vol. 15, No. 11, p.e0240915.

Min, J., Gao, Z., Wang, L. and Zhang, A. (2024) '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.

Moon, H.S., Orr, G. and Jeon, M. (2023) 'Hand tracking with vibrotactile feedback enhanced presence, engagement, usability, and performance in a virtual reality rhythm game', *International Journal of Human-Computer Interaction*, Vol. 39, No. 14, pp.2840–2851.

Nieto-del-Amor, F., Beskhani, R., Ye-Lin, Y., Garcia-Casado, J., Diaz-Martinez, A., Monfort-Ortiz, R., Diago-Almela, V.J., Hao, D. and Prats-Boluda, G. (2021) 'Assessment of dispersion and bubble entropy measures for enhancing preterm birth prediction based on electrohysterographic signals', *Sensors*, Vol. 21, No. 18, p.6071.

Qiu, L., Li, S. and Sung, Y. (2021) 'DBTMPE: Deep bidirectional transformers-based masked predictive encoder approach for music genre classification', *Mathematics*, Vol. 9, No. 5, p.530.

Ren, L., Yu, G., Wang, J., Liu, L., Domeniconi, C. and Zhang, X. (2022) 'A diversified attention model for interpretable multiple clusterings', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 9, pp.8852–8864.

Śmigiel, S., Pałczyński, K. and Ledziński, D. (2021) 'ECG signal classification using deep learning techniques based on the PTB-XL dataset', *Entropy*, Vol. 23, No. 9, p.1121.

Subba, T. and Chingtham, T. (2024) 'Comparative analysis of machine learning algorithms with advanced feature extraction for ECG signal classification', *IEEE Access*, Vol. 12, pp.57727–57740.

Torres-Soto, J. and Ashley, E.A. (2020) 'Multi-task deep learning for cardiac rhythm detection in wearable devices', *NPJ Digital Medicine*, Vol. 3, No. 1, p.116.

Wahid, A., Yahya, M., Breslin, J.G. and Intizar, M.A. (2023) 'Self-attention transformer-based architecture for remaining useful life estimation of complex machines', *Procedia Computer Science*, Vol. 217, pp.456–464.

Wei, W., Wang, Z., Mao, X., Zhou, G., Zhou, P. and Jiang, S. (2021) 'Position-aware self-attention based neural sequence labeling', *Pattern Recognition*, Vol. 110, p.107636.

Xiong, Q., Zhang, J., Wang, P., Liu, D. and Gao, R.X. (2020) 'Transferable two-stream convolutional neural network for human action recognition', *Journal of Manufacturing Systems*, Vol. 56, pp.605–614.

Zhao, F., Zhang, C. and Geng, B. (2024) 'Deep multimodal data fusion', *ACM Computing Surveys*, Vol. 56, No. 9, pp.1–36.