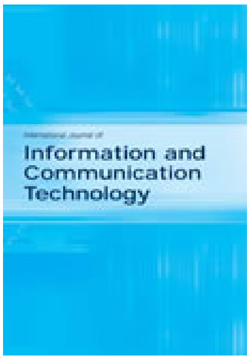
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Design and practice of artificial intelligence-driven piano improvisation accompaniment teaching system introduction

Xiang Wei

DOI: 10.1504/IJICT.2025.10075323

Article History:

Received:	28 August 2025
Last revised:	14 September 2025
Accepted:	16 September 2025
Published online:	15 January 2026

Copyright © 2025 Inderscience Enterprises Ltd.

Design and practice of artificial intelligence-driven piano improvisation accompaniment teaching system introduction

Xiang Wei

College of Architecture and Arts,
Taiyuan University of Technology,
Taiyuan, Shanxi, 030024, China
Email: weixiang@tyut.edu.cn

Abstract: In this study, we provide a system that shows students how to play piano with improvisation and accompaniment using cloud computing, deep learning, and CNN. Automatic evaluation of performance aspects, such as pitch, timbre, articulation, rhythm, and dynamics, is one way the suggested approach enhances piano lessons. Applying a hybrid approach that combines a matched filter with a rapid guided filter optimises preprocessing for music feature extraction. To further improve the accuracy of piano performance analysis, attention-induced multi-head CNNs and perceptual evaluation datasets are employed. In adaptive and remote learning settings, the technique shows better dependability and scalability. The model successfully integrates visual and aural methods of teaching piano, supports multilevel perceptual feature analysis, by providing a novel framework that enhances learning outcomes, enables tailored instruction, and adapts to the diverse needs of learners, this research contributes to the expanding field of intelligent music education.

Keywords: artificial intelligence; AI; piano teaching; improvisation accompaniment; convolutional neural network; CNN; deep learning; DL; cloud computing; perceptual features; music education technology.

Reference to this paper should be made as follows: Wei, X. (2025) 'Design and practice of artificial intelligence-driven piano improvisation accompaniment teaching system introduction', *Int. J. Information and Communication Technology*, Vol. 26, No. 52, pp.56–74.

Biographical notes: Xiang Wei is a researcher at the College of Architecture and Arts, Taiyuan University of Technology, Taiyuan, Shanxi, China. His research area is arts education.

1 Introduction

Engineering, AI, internet technology, music, and numerous other fields have all begun to incorporate computer music technology into their work in recent years. Composers are motivated to create music in unique ways when they use computers to freely compose music with the help of algorithms in songwriting programs (Li, 2022). Additionally, computer music technology can provide a pervasive and inexpensive music tutoring service. An automated system that can assess a pianist's performance on several factors

Copyright © The Author(s) 2025. Published by Inderscience Publishers Ltd. This is an Open Access Article distributed under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

(e.g., rhythm, articulation, expressiveness, timbre, pitch, and chords) is our goal in doing this research. If the system can determine the user's current skill level and provide them with timely feedback on how to improve, it would be beneficial for piano students (Alzubaidi et al., 2021). It is essential to consider a wide range of talents when evaluating a student's progress in piano lessons, especially for younger children. Consequently, there are issues in education and learning that can be addressed, and abilities in face-to-face instruction at various levels can be enhanced through the use of computer-based techniques (Phanichraksaphong and Tsai, 2023). We include timbre- and pitch-based evaluation tools, as playing the piano requires a multifaceted set of skills, including control over volume and dynamics, as well as rhythms, techniques, body language, and facial expressions.

Due to its demanding nature, playing the piano is an excellent way to develop stronger hand-eye and motor skills. When playing the piano, it's essential to use both your left and right hands. But you can't rely on one hand to play the melody or rhythms alone; for example, you may play the melody with your right hand and the accompaniment with your left, giving the impression that both hands are acting separately. The independence of the hands provides the pianist with greater leeway to express themselves while playing. In addition, you can use both of your feet to press down on the pedals at the same time. There has been a diverse trend in the development of computer-assisted composition over the last half-century. Artificial neural networks, genetic algorithms, music grammar rules, and other similar techniques are the mainstays of automatic composition. While these approaches can address some of the requirements of autonomous composition, they are not without their flaws (Peñalver Vilar and Valles Grau, 2020). Take a recurrent neural network music, for example. It lacks overall musical coherence and attempts to correct the melody and harmony using a genetic algorithm, only to end up creating meaningless local optimal regions in the harmonic search space problem.

The inability of computer-assisted composition to keep up with the ever-evolving nature of musical materials is currently its biggest challenge (Stün and Ozer, 2020). From one angle, music is simply a combination of various musical parts, and computers excel at mathematical calculations. Contrarily, computers lack human emotions and thought processes, and music is an art form. Therefore, computer-assisted composition necessitates more assistance from AI technology, in addition to more diverse programs. As it stands, the music programs at public and private schools make excellent use of technological resources. The compatibility of their music classroom setting with various social cultures is also undergoing minor alterations as a result of the shift in communication style between instructors and students (Li, 2020). School administrators can, on the one hand, utilise big data analysis to pinpoint inefficiencies in the current instructional method and refine it. As an alternative, innovative education has introduced new approaches to teaching and learning, including the intelligent piano, and various forms of music learning software have altered the way people study music (Liu and Huang, 2021). There is potential for intelligent piano instruction to leverage deep learning (DL). Take, for example, DL-based automatic music transcription. To provide an unbiased justification for the accuracy of performance, let students quickly identify their mistakes, and improve learning efficiency, it is beneficial to compare the played music to a standard score. Piano grading tests and computer-assisted piano instruction are two areas that could benefit from this technology. The study's first section includes an analysis of the literature on intelligent music instruction (Li, 2022). Using the cognitive and motor growth of preschoolers as a lens, the second segment delves more into the features and capabilities of intelligent pianos.

This serves as the basis for building a convolutional neural network (CNN) that detects the onset of piano notes. Additionally, to determine people's opinions on the educational and popular effects of intelligent pianos, we surveyed parents and preschoolers. This can be used as a realistic basis for intelligent piano promotion and instruction. The poll's findings and the CNN model's performance are covered in the fourth part. The structure of this manuscript is as follows: Section 1 introduces the research background and motivation. Section 2 reviews relevant literature on AI-driven piano design and teaching practices. Section 3 describes the study's methodology and the improvisation accompaniment teaching system. Section 4 reports the experimental findings and discusses their implications. Finally, Section 5 summarises the conclusions and outlines directions for future research.

1.1 Paper contribution

Incorporating cloud computing, artificial intelligence (AI), and DL into a unified framework for training and evaluation, this study contributes to the growing body of research in intelligent piano education. At its core, the program is an AI-driven improvisation and piano accompaniment system that can assess students' progress along multiple dimensions and tailor its feedback to each individual's strengths and areas for improvement. The study employs a hybrid preprocessing strategy that combines matched filters with quick guided filters to enhance the accuracy of music feature extraction and ensure reliable assessment across rhythm, dynamics, articulation, and pitch. Thirdly, attention-induced multi-head CNNs integrate visual and auditory teaching methods by enabling a more comprehensive investigation of perceptual performance traits. The research concludes that the system can be easily scaled, made accessible, and adapted to various learning contexts (such as remote and tailored education) by implementing it within a cloud -based architecture. Taken as a whole, these papers provide light on the state of intelligent music education and propose new ways to teach piano with the help of technology, both theoretically and practically.

2 Related work

2.1 Conflicts in teaching approaches to early piano education

The master-apprentice method, in which a student learns an instrument and its repertoire by emulating the actions and intonation of a more experienced player, is one of the inventive approaches to music education that McPherson and Gabriel's son recalled. Nonetheless, the majority of contemporary method books use a visual approach that connects the fingers to notation rather than sound, thereby enhancing the mathematical correlations between scale degrees. They frequently divide the process of learning technical competence from learning to play actual music, prioritising note identification and theoretical concepts over gaining perceptual comprehension. Teachers may use visual aids, such as fingerings, letter names, and hand posture, to help students learn the C scale, for instance. As an additional illustration, consider how pre-staff notation teaches high and low registers using visual connections rather than auditory cues (Kan, 2022).

Early piano instruction currently centres on visual musical notation literacy, according to Bunting, Williams, and Arshinova.

Piano method books cover the fundamentals of the instrument, and it looks like music reproduction is the primary focus of early piano lessons. Concepts like black and white keys, instructions, geography, letter names, and number and letter ranges are better understood by students when they see them illustrated in method books. They also help students with concepts like hand and finger positioning and recognition. Good tools for teaching piano to children typically incorporate note reading into the very first lesson, along with the fundamentals of music theory and the instrument's mechanics, and utilise eye-catching visuals to pique the interest of young learners (Pang, 2024a). The fact that these resources are attractive to piano teachers is not surprising. Nevertheless, to achieve the objective of music reproduction, this encourages a theory-driven understanding of music. However, this does not rule out the use of auditory methods in the early stages of piano training. On the contrary, such events are often overlooked. Books such as *Music Little Mozart's*: Books like *Music Lessons: Book 1 and Prep Course for Young Beginners*: Lesson Book, Level A, both have short musical parts that teach kids how to incorporate their voices and bodies into singing and percussion.

2.2 Application in intelligent electronic musical instruments

Technological advancements in AI over the last several years have enabled electronic musical instruments to become more sophisticated, individualised, and intelligent, ushering in a new era. In addition to storing a wide variety of musical instrument timbres, the intelligent electronic instrument can also effectively combine timbres, allowing for the execution of timbres in response to a variety of action instructions (Zheng, 2022). Classical musical instruments clearly lack the functionality necessary to accomplish this task. Because of these advantages, intelligent electronic instruments are slowly but surely making their way into music classrooms. A new way of teaching music has emerged with the advent of intelligent electronic musical instruments. Now more than ever, a solo musician can inspire new ideas by experimenting with different combinations of powerful sounds (Zhang, 2023). Students of music practice greatly benefit from music, and they also achieve a greater level of instruction as a result. In today's world, where science and technology are advancing at a rapid pace, AI is becoming increasingly significant, along with digital technology, online performances of electronic music, and collaborative research on wireless networks. A system for making electronic music was developed in the area of AI.

Thanks to this technology, online schooling can now utilise both wireless networks and electronic music (Yu and Ma, 2023). A new electronic musical instrument has emerged due to technological advancements in computer sensor networks, intelligent algorithms, and wireless networks. We can verify the degree of alignment between the AI electronic music course materials and the objectives of online intelligent matching and online education by running a simulation experiment. The information about the sound is subsequently transformed into visual and auditory patterns using the oscilloscope.

2.3 Adaptive piano accompaniment

Adaptive piano accompaniment generative adversarial networks are an area where our method excels, surpassing both competing GANs and innovative hybrid methods. The data processing capabilities of this tool are second to none when compared to mixing technology, thanks to its utilisation of cutting-edge technology for information mining and the provision of high-quality data to the model, which allows for considerably more flexible accompaniment and melody generation (Kale and Altun, 2024). Due to issues with gradient vanishing and other model-level problems common to typical GANs, this tool utilises adaptive ensemble methods to enhance training stability and employs Gaussian mixture models to generate a diverse range of coherent accompaniments. This technology utilises sentiment analysis and other techniques to make the melody and accompaniment emotionally engaging, which distinguishes it from other GAN arts that struggle with musical understanding and expression (Karamatlı, 2024). In a nutshell, this instrument spearheads the field's development and accomplishes advances in numerous dimensions. Environments can be changed via adaptive integration technology. While block-based integration manages data in separate pieces, online integration handles training instances without requiring storage.

Most adaptive weighted integration techniques use SEA or AWE when working with data blocks (Zhou, 2025). While this kind of algorithm excels at handling gradual drifts in concepts, it is notoriously slow to react to abrupt changes. Current ensemble techniques that utilise blocks to train classifiers employ recently tagged data to inform the categorisation of unlabelled data. On the other hand, the ensemble model might not give reliable results if there is idea drift in the unlabelled data. Unlabelled data can include valuable insights that current algorithms fail to capture. Consequently, they are unable to adapt to the present environment by monitoring concept drift over time or by quickly assigning appropriate weights to component classifiers.

3 Proposed methodology

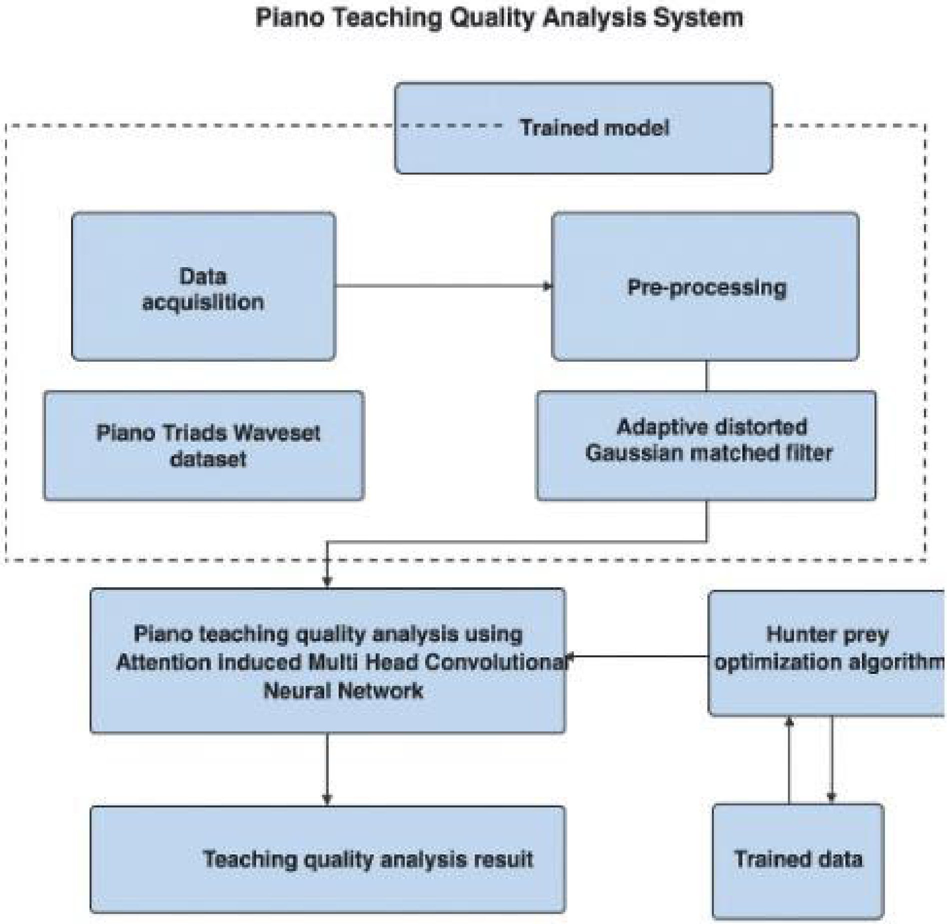
Using cloud computing, AI, and machine learning, this study presents RPT-AIMCNN-HPO, a system for distant piano instruction. Figure 1 illustrates the block diagram of the RPT-AIMCNN-HPO method. Using AIMCNN in the cloud, the following is a detailed example of remote piano instruction (Song, 2024).

3.1 Trained model

Twenty-five human pianists and two 'score' performances make up the 1,202 musical portions that make up PercePiano, which has 12,652 annotations. A total of 6,244 annotations, 10,219 annotations, and 1,809 annotations make up these portions. It was possible to compare the two 'score' performances with the human performances because they were taken directly from the original MusicXML score 4,647. Choose from 'Score' and 'Score2'. In contrast to the mechanical quality of the latter, the former makes greater use of musical notations (such as legato and dynamics) to mimic human performance. A total of 53 separate annotators, each rating 19 distinct labels, have evaluated the annotations. To be more precise, the following Schubert compositions have a combined total of 4,076 annotations: D.960, mv2 (2nd movement), D.960, mv3 (3rd movement), D.935, with 624 annotations, and Wo O.80 by Beethoven, which has 1,244 annotations. With a standard variation of 3.62, each performance segment typically has 10.52 annotations. The average and standard deviation of annotator ratings across all 19 criteria for each performance are displayed in Table 2. The pieces' names span from WoO.80 to

D.950 mv2. An auxiliary file showing the mean and standard deviation is Figure S2 (Park et al., 2024).

Figure 1 Schematic representation of the RPT-AIMCNN-HPO method (see online version for colours)

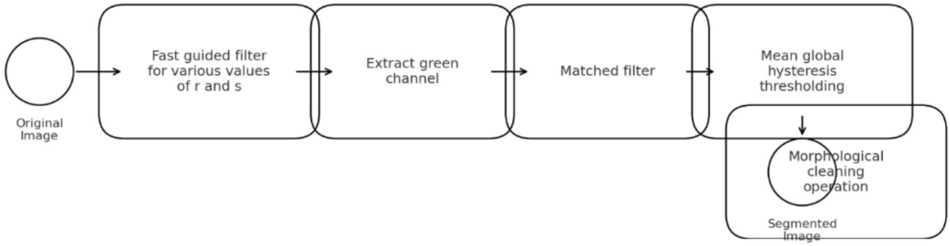

Piano Teaching Quality Analysis System
Trained model
Data acquisition
Pre-processing
Piano Triads Waveset dataset
Adaptive distorted Gaussian matched filter
Piano teaching quality analysis using Attention induced Multi Head Convolutional Neural Network
Hunter prey optimization algorithm
Teaching quality analysis result
Trained data

‘I do not know’ was appended to 921 of the 12,652 annotations, or 7.3% of the total. You can find the exact total down below. We provide thorough statistics and data quality checks for each musical composition in supplementary notes A. One popular descriptive statistic for quantitative measures is the intraclass correlation coefficient (ICC), which shows how similar different units are within the same group. In order to investigate the level of agreement between annotators, we sort ICC evaluation models by the data and annotators that were employed [ICC (1, 1) and ICC (1, k)]. When a separate group of k randomly chosen annotators measures each subject, for instance, one-way random evaluation is employed (ICC, 1, k). An random one-way assessment was determined to be suitable49 after employing distinct sets of randomly selected annotators for each section. To determine the reliability of the averages and individual evaluations, we calculated the ICC (1, 1) and ICC (1, k) values for each label (Pang, 2024b). When examining average reliability, the ICC is ‘excellent’ (1, k), according to Table 2; however, when considering single-measure dependability, it is ‘poor’ (1, 1). These results suggest that people’s subjective views on music may not be comprehensive, but when considered collectively, they tend to converge towards a more widely accepted understanding of music.

3.2 Adaptive distorted Gaussian matched filter for pre-processing

In most cases, greyscale retinal pictures do not clearly show small blood vessels. Because of the lack of contrast in local intensity, vessel segmentation is severely limited. The intensity of the vessel’s width, which encompasses its borders, differs considerably among photographs. Equally mixed with Gaussian noise are little vessels. Consequently, the majority of the methods proposed in the literature for precisely identifying vessels have been unsuccessful. Due to this limitation, vessel segmentation is a challenging task. It is clear from the equation sets and the quick guided filter description that pixels in regions with high variance will keep their values, while pixels in areas with even variance will have their values smoothed out by nearby pixels. Therefore, with a frequency defined by an averaging method, very few fine features in the virtually flat portions are smoothed away (Dash, 2022). One easy and effective way to remove vessels is with a matching filter. A matching filter can detect edges on vessels as well as those outside of them. On the other hand, a guided filter is an operator that performs better at the edges, exhibiting both smoothing and preserving qualities. Based on these characteristics, combining a matched filter with a rapid guided filter in a single model will improve vessels and allow for precise vessel extraction. Figure 2 depicts the three stages of the proposed procedure.

Figure 2 Proposed method’s schematic diagram


Original Image
Fast guided filter for various values of r and s
Extract green channel
Matched filter
Mean global hysteresis thresholding
Morphological cleaning operation
Segmented Image

3.2.1 Matched filter

It is possible to identify blood vessels using a Gaussian matching filter when the vasculature’s grey-level profile approaches a Gaussian-shaped curve. Below is a summary of the matching filter, and you can find its specifics in the documentation. Here is a description of the matching filter that uses the Gaussian kernel function:

$$P(m, n) = \exp\left(-m^2 \big/ 2\sigma^2\right) \quad \forall \,|\, n \,|\leq \frac{L}{2}, |\, m \,|\leq t.3, \tag{1}$$

Where the matched filter is defined as

$$Q(m,n)=\left[\frac{-1}{\sqrt{2\pi\sigma}}\right]\times\exp\left(\frac{-m^2}{2\sigma^2}\right)-m \quad \forall\,|\,n\,|\leq\frac{L}{2},|\,m\,|\geq t.3, \tag{2}$$

$$B=\frac{\int_{-ts}^{ts}\left[\frac{-1}{\sqrt{2\pi\sigma}}\right]\times\exp\left(\frac{-m^2}{2\sigma^2}\right)dm}{2ts} \tag{3}$$

The vessel segment length, denoted as *L*, the intensity outline spreading, denoted as *σ*, and a constant *t*, fixed at 3, constitute smooth noise. The vessel identification process involves maximising the filter bank's response by rotating the kernel *P*(*m*, *n*) in different orientations. Twelve kernels rotated at 15-degree intervals are sufficient to accurately identify the vessels. In a Gaussian curve where the signals are infinitely long, the two-sided tails are cut off at *u* =± 3*σ*, and N is represented as

$$N=\left\{(u,v), VuV\leq 3\sigma, V\vartheta V\leq {}^{L}\!/_{2}\right\} \tag{4}$$

The weights in the kernels *i* (where *i* ranges from 1 to 12, the total number of kernels) are defined by.

$$p_t(m,n)=-\exp\left(\frac{-u^2}{2\sigma^2}\right) \quad \forall Z_1\in N \tag{5}$$

The following is the formula for determining the kernel mean value when *A* is a set of points in *N*:

$$s_t=\sum_{Z_1\in N} {}^{P_1(m,n)}\!/_{A} \tag{6}$$

Hence, this is the convolution mask:

$$P(m,n)=p_f(m,n)-S_f \quad \forall Z_f\in N \tag{7}$$

3.2.2 Fast guided filter

Although it performs better near the edges, a directed filter is essentially a special case of a bilateral filter. Theoretically, a directed filter might interact with the Laplacian matrix. Moreover, guided filters can utilise structures to enhance the quality of the output image, which is not the case with regular smoothing operators. The computational complexity is independent of the filter's kernel size because the guided filter uses a fast and non-approximate linear-time technique. Noise reduction, HDR compression, enhancement, haze removal, and joint upsampling are just a few of the many uses for guided filters in computer vision, computer graphics, and computer science. A guided filter takes input images *I*, uses guidance images *P*, and produces filtered output images *Q* using a basic linear model. Knowing that pixel *k* is the centre of window m_k, it is necessary to assert that the linear transform is *Q* of *P*.

See the kernel and guided filter definitions below:

$$Q_t=C_k l_t+d_k, \quad \forall_1\in\in_k \tag{8}$$

(*ck*, *dk*) are linear coefficients that are almost constant, and the *wk* square window has an index of *k* and a radius of *r*.

The output is the minimum reconstruction error between *P* and *Q* as determined by equation (8) for input image *P*

$$C_k = \frac{\frac{1}{|w|}\sum_{t\in w} I_t P_t - \mu_k \bar{P}_k}{\sigma_k^2 + \varepsilon} \tag{9}$$

$$d_k = C_k \mu_k$$

The degree of smoothness is controlled by the regularisation parameter *ε*, where *μk* represents the mean and *σk* stands for the variance of *I* in the window. Following the computation of (*ck*, *dk*) for each image patch *wk*, the following steps are taken to determine the filter output:

$$Q_t = \frac{1}{|w|}\sum_{k:1w_k}\left(C_k p_1 + d_k\right) \tag{10}$$

$$Q_t = \bar{C}_1 I_t + \bar{d}_t \tag{11}$$

where $c_i = 1|w|\sum k \in wick$ and $d_i = 1|w|\sum k \in widk$ are the mean values of the coefficients for all *i*-centred windows. The first algorithm shows the procedures that the guided filter follows. When approaching *O*(*N*) time, *Zmean* represents the mean filter with considerable variability. While conventional guided filters rely heavily on the guiding image, they struggle to achieve fast computation when denoising images, which is why you should use a rapid guided filter. The time complexity for a subsampling ratio s can be reduced from *O*(*N*) to *O*(*Ny*2/) using a fast guided filter. A fast guided filter outperforms the standard by a factor of ten in many cases, all without sacrificing performance.

Algorithm 1 Algorithm for guided filter

```
Input parameters: $A$ is the input filtering image, $P$ is the guidance image, $r$ is the radius, and $\varepsilon$ is the regularisation.
Output parameter: $Q$ is the filtering output.
1  $meanP = fmean(P)$
       $meanA = fmean(A)$
       $corrP = fmean(P \times P)$
       $corrPA = fmean(P \times A)$
2  $varP = corrP - meanP * meanP$
       $cosPA = corrPA - meanP * meanA$
3  $x = covPa./(varP + \in)$
       $y = meanA - x. * meanP$
4  $meanx = fmean(x)$
       $meany = fmean(y)$
5  $Q = meanx. * \underline{\mathrm{P}} + meany$
```

3.2.3 Preprocessing

The utilisation of retinal fundus images allowed for the automatic detection of eye disorders in fundus images. The most challenging part of interpreting a fundus image, though, is dealing with the image corruption, which can happen for a number of reasons. The quality of a fundus image is diminished due to a cataract in a human lens, similar to how a hazy camera lens reduces the clarity of a photograph. The contents and properties of the photos are altered based on fundus images from various clinical circumstances found in different databases. As a result, it is essential to enhance the overall image quality during the pre-processing steps. Combining a directed filter with a matched filter is an innovative strategy that can improve retinal vascular performance measures. Improving the image's overall quality was the first stage in using the rapid guided filter. Because the green component makes retinal arteries more visible and contrasty than the blue and red ones, it was the next step in the vessel extraction process to apply it exclusively to the matching filter.

3.3 Analysis of piano teaching effectiveness using an attention-driven multi-head CNN

Here, we examined the ConvNet-MPE baseline model, a CNN-based MPE model. However, local polyphony estimation (LPE) was performed using a number of CNN models that were trained on the new feature representations. Differences in kernel shape and number of pooling layers are the primary distinguishing features of these models. The following sections provide detailed descriptions of each model type, summarised in Table 1 (Dash, 2022).

Table 1 Synopsis of neural network model structures, objectives, feature representations, convolutional block kernel shapes, class numbers, and model parameters

Model	*Task*	*Feature representation*	*Kernel shape*	*Number of classes*	*Number of parameters*
ConvNet-MPE	MPE	HR-CQT	(3, 3)	88	2,158,000
ConvNet-LPE	LPE	HR-CQT	(3, 3)	3 / 6 / 13	2,114,000–2,119,000
CQT	LPE	HR-CQT	(1 / 3 / 5, 24)	3 / 6 / 13	79,000–287,000
CQT	LPE	LR-CQT	(3, 24)	3 / 6 / 13	112,000–161,000
F-CQT	LPE	F-CQT	(4, 3 / 6)	3 / 6 / 13	47,000–107,000
F-CQT 3D	LPE	F-CQT	(3 / 5, 4, 3 / 6)	3 / 6 / 13	221,000–394,000

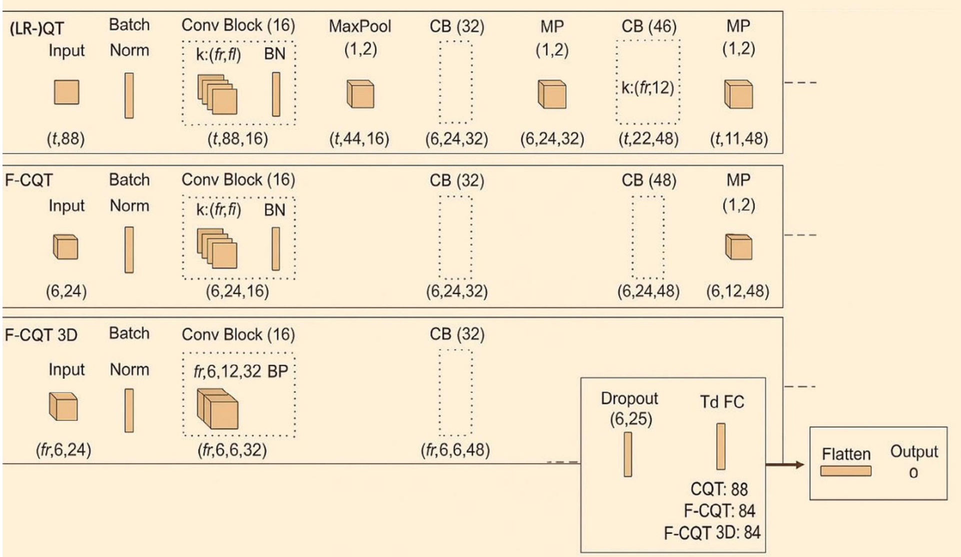
3.3.1 MPE model: ConvNet-MPE

All comparisons in MPE are based on the 'ConvNet' architecture, which was initially developed by. Our reimplementation produced the same results as the original publication, as demonstrated by the MPE results reported in. Specifically, the ConvNet-MPE is composed of three CBs. Batch normalisation, ReLU activation, dropout, and convolutional layers are all components of each CB. Only during training was the dropout clause invoked. With momentum-based batch normalisation, both of the first CBs used a kernel size of (3, 3) and 32 filters. A 0.1 momentum was setup. In the context of convolutional approaches, 'valid padding' refers to the absence of zero-padding. A max-pooling layer with a pool size of (1, 2) and a dropout layer with a minimal dropout of 0.25 followed these two CBs. The most current CB used 64 filters instead of 32 and a reduced kernel size of (1, 3). Two dense layers, one with 512 units and the other with 88 units, were employed following the three CBs. The sigmoid activation function was helpful in this situation. Additionally, a dropout probability of 0.5 was used between the two dense layers. There are no bias terms in any of the model's computational levels. Training the ConvNet-MPE involved reducing the binary cross-entropy loss function and HR-CQT.

3.3.2 LPE models

Figure 3 shows every LPE model. All the LPE models considered are derivatives of the same basic design, in contrast to ConvNet-LPE. Everything about the last dense layer stack, including input batch normalisation (BN), indicates that each convolutional layer (CB) has the same number of filters. There are three CBs in total.

Figure 3 Specifically, the three CNN models utilised in the LPE studies are described in detail, including the output tensor forms for each layer (see online version for colours)



3.3.3 Class, fifth, frame, feature, kernel, octave, and time are all abbreviations. Included are additional specifications for the parameters

A convolutional layer, a BN, and an activation function for a rectified linear unit (ReLU) make up each CB. A probability of 0.25 was used to apply dropout after the last CB. We then used the same number of units as the bins in the original feature representation to train a bias-free fully connected (FC) layer with a tanh activation function. After that, another FC with the same amount of units as the current polyphonic scenario and Softmax activation followed. Default stride sizes were one, and all CBs utilised the same padding. Lastly, the categorical cross- entropy loss function was used for training all LPE models except ConvNet-LPE.

3.3.4 ConvNet-LPE

To train the model for the LPE task, we used a softmax activation function and changed the number of units in the ConvNet-MPE's output layer to 3, 6, or 13 (as indicated).

We were able to determine if the overall ConvNet design was effective for LPE in this manner. Like the ConvNet- MPE, this model kept the exact input feature representation, optimiser, and learning rate settings.

3.3.5 CQT model

Based on the batch size, the CQT model received CQTgram parts as input. These chunks may come from the LR- CQTgram (with feature dimensions of 88) or the HR-CQTgram (with feature dimensions of 264), depending on the specific model. *Za* is the number of frames it includes, and *Zt* is the number of surrounding CQT-features (bins), which we used as the kernel size for the first two CBs. If we want to see how their differences play out, we can compare the kernel's and the F-CQT's training sets, which both cover the same amount of notes but with different bin sizes: 12 for the kernel and 24 for the F-CQT. The pooling layers gradually downsampled data from the feature dimension. In the latest CB, we had to adjust the kernel to 12 bins to prevent over-patching and reduce the parameter count.

3.3.6 F-CQT model

Data from specific frames of the F-CQT model were input into it. A single frame had dimensions of 6 × 24 due to the two-dimensional structure imposed by the F-CQT arrangement, which utilised 12 bins per octave and 7 octaves in total. In the original CB, the kernel size was (*o*, *ZT*), where *Zi* is the octave number and *Zi* is the fifth number. Since it might cover 4 octaves and three fifths, the proposed kernel size is (4, 3). The benefit of covering harmonics within an octave is achieved by using an F-CQT convolutional kernel with fewer parameters. The F-CQT model avoided all intermediate pooling stages that follow each CB due to its smaller kernel size and reduced number of model parameters, except for the last max-pooling operation over the fifth-related tensor dimension.

3.3.7 F-CQT 3D model

The F-CQT model and its 3D counterpart are structurally identical. Even though both the CQT and F-CQT 3D models utilise temporal context, the former employs convolution kernels that span multiple frames, rather than just one. We created a three-dimensional input feature by combining numerous successive F-CQT frames, enabling the F-CQT 3D model to receive data. The input was scanned in the CBs using a 3D kernel that resembled (*Za*, *o*, *Zi*). We decreased the feature representation during the forward pass because this method could lead to a much greater parameter count: A single row of octaves, or the fifth dimension, was used to add pooling layers following the first two CB.

4 Results

We investigated the relationships between several perceptual features of piano playing and various hierarchical levels (note, voice, beat, measure), from the most basic to the most complex. We trained and evaluated each succeeding model using outputs from different hierarchical levels, starting with the Bi-LSTM baseline. A randomly split dataset was utilised to eliminate the possibility of bias towards particular works or performers. To start, the addition of hierarchical information has a greater impact on lower-level perceptual qualities, including time, articulation, and timbre, as seen in the line chart in Figure 4. Given the direct association between hierarchical information and lower features, this is in line with expectations. Secondly, the bar chart clearly demonstrates that perceptual feature performance is enhanced across the board when hierarchies are used at the note, voice, and rhythm levels. The effects across several measures are inconsistent. Features at low, mid-low, and high levels perform better, but features at mid-high levels do not see significant improvement. 'Dynamic' and 'music creating' are best conveyed and perceived by paying close attention to the expression levels along each voice; in the medium-high range, beat-level timing adjustments are more significant than measure-level ones. For these reasons, we have arrived at these conclusions.

4.1 Necessity of RA metric

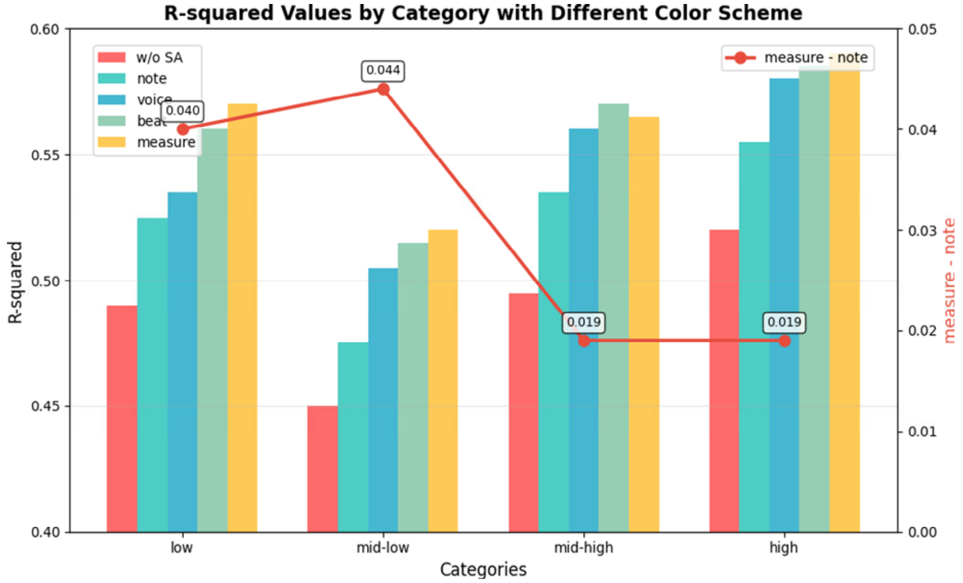
A noticeable pattern emerged in the evaluation results: when applied to data with high levels of annotator disagreement, the model's performance dropped significantly. The interaction between the Bi-LSTM+SA+HAN model and performance metrics (MSE and standard deviation) is seen in Figure 5. The figure illustrates that this trend emerges, highlighting the importance of exercising caution when evaluating the model's performance, particularly when considering subjective perceptual assessments. This finding sheds light on the reasoning behind using the RA metric. Low standard deviations, which show strong agreement among annotators, are penalised by the RA metric. To put it simply, the metric emphasises objectivity by penalising cases where annotators converge heavily, drawing attention to possible errors in the model's predictions. The converse is also true; instances with larger standard deviations show that it allows for a greater degree of subjectivity. To comprehend the model's efficacy across different degrees of annotator consensus, one must employ this intricate metric, which takes into consideration the subjective components of the annotated data.

We found that the proposed method, which combines matched filters and rapid guided filters, outperformed the alternatives and produced superior results. A distinction was discernible in the healthy gradient vector field between the pixels representing the vessels and the diseased tissues. Improved specificity but unimpressive sensitivity resulted from labelling as 'non-vessel pixels' pixels that were brighter than their neighbours. Table 2 shows how the proposed strategy stacks up against other cutting-edge approaches. Research shows that most comparisons were made with pre-existing matching filters because that was the intended goal of the suggested method. The proposed approach yields performance matrices that are significantly superior to those of state-of- the-art methodologies.

Table 2 Examining the proposed approach's performance

Approach	*Year*	*Sn*		*Spc*		*Ac*	
		DRV	*CDB*	*DRV*	*CDB*	*DRV*	*CDB*
Dash	2020	0.7203	0.6454	0.9871	0.9799	0.9581	0.9609
Dash and Senapati	2020	0.7403	--	0.9905	--	0.9661	--
AlSaeed	2020	0.6312	--	0.9817	--	0.9353	---
Memari	2019	0.761	0.738	0.981	0.968	0.961	0.93
Subudhi	2016	0.3451	--	0.9716	--	0.911	--
Sreejini and Govindan	2015	0.7132	--	0.9866	--	0.9633	--

Figure 4 Measuring the model's effectiveness over four assessment levels: low, mid-low, mid-high, and high, with different music hierarchies: note, voice, beat, and measure (see online version for colours)



Note: You can see how well each model did on its own in the bar chart, and how far off note-based and measure-based predictions were in the line chart.

The original photographs of retinas 2 and 4, located in the DRV database, and retina 1 in the CDB database, are displayed in Figures 6(a), 7(a), and 8(a), respectively. These are the actual images of retinas 2 and 4, collected from the DRIVE and CHASE databases, respectively: Figures 6(b), 7(b), and 8(b) provide further details. Retinas 2 and 4 from the DRV database and retina 1 from the CDB database were used to construct the vessel-extracted image, which is shown in Figures 6(c), 7(c), and 8(c). Figures 6(d)–6(h), 7(d)–7(h), and 8(d)–8(h) show the recovered vessels that were processed using the suggested procedure. The images were captured using DRV database Rips 2 and 4, and CDB database Retina 1. The quick guided filter was adjusted for each setting. Statistical analysis reveals that the suggested method outperforms the original matched filter in detecting thin vessels while producing fewer false positives.

Figure 5 The correlation between standard deviation and MSE performance, indicating model underperformance with low annotator agreement (see online version for colours)

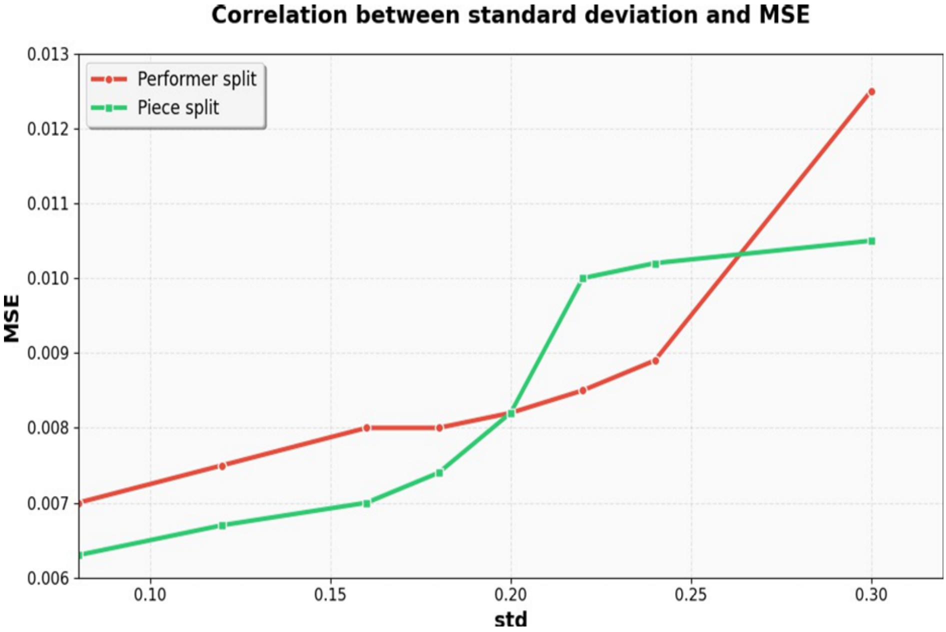
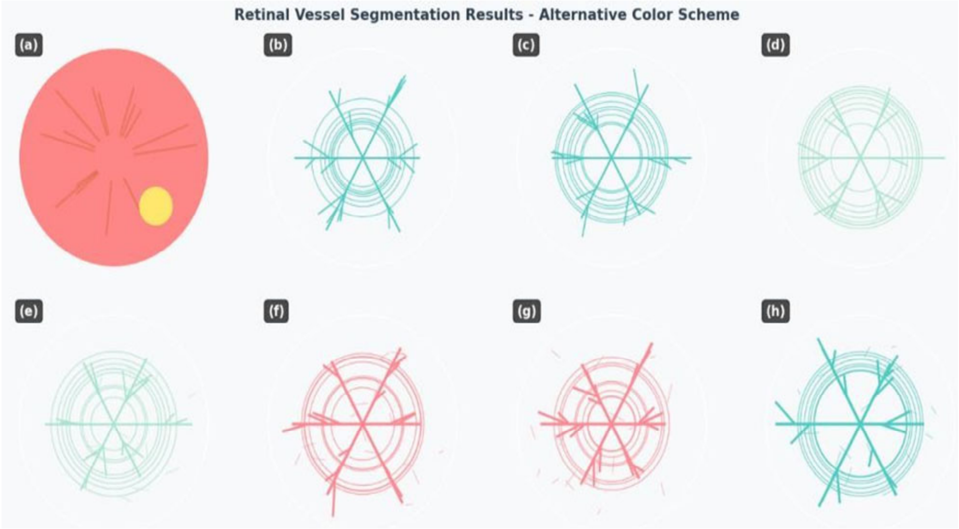


Figure 6 Matched filter technique and fast guided filter for retina 2 dataset (see online version for colours)


Retinal Vessel Segmentation Results - Alternative Color Scheme
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

Segmentation based on multiple parameter values: The following images serve as examples of parameter combinations: (a) the beginning point image, (b) a picture of the ground truth, (c) the original matching filter, and (d)–(h) the images that have been segmented for the following values of s and r: 1, 6, 2, 10, 3, 18, 4, 28, and 5, 25, respectively.

Figure 7 Four retina images were extracted from the DRV dataset using the fast guided filter and matched filter methods with varying parameter values, these images include (a) the original image, (b) the ground truth image, (c) the original matched filter, and (d)–(h) separated images for various parameter combinations, including s = 1, r = 6, s = 2, r = 10, s = 3, r = 18, s = 4, r = 28, and s = 5 (see online version for colours)

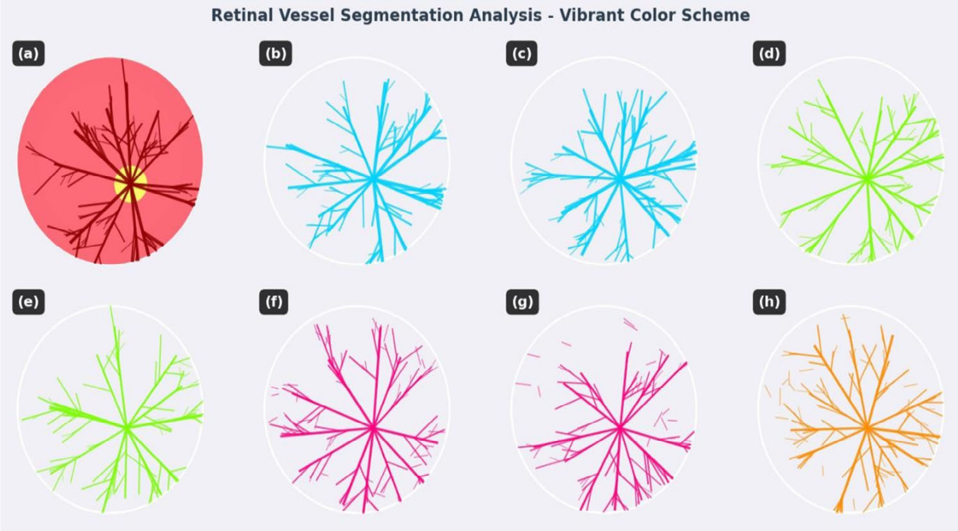
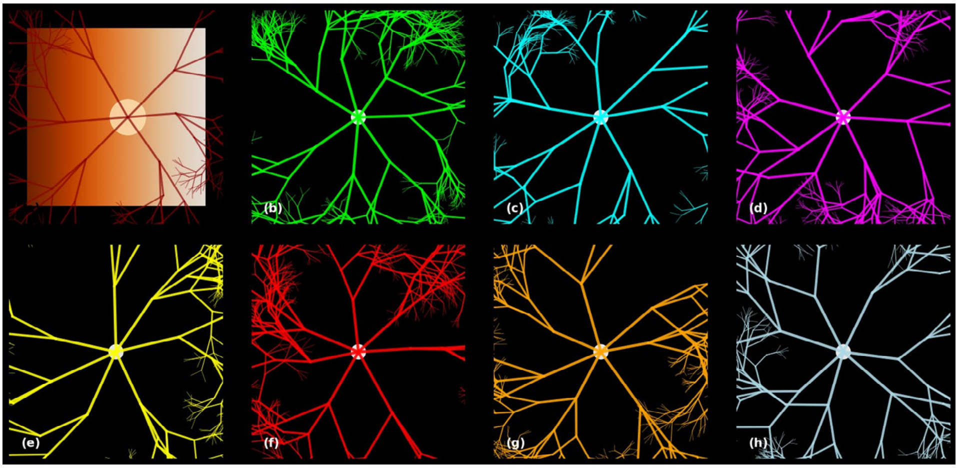


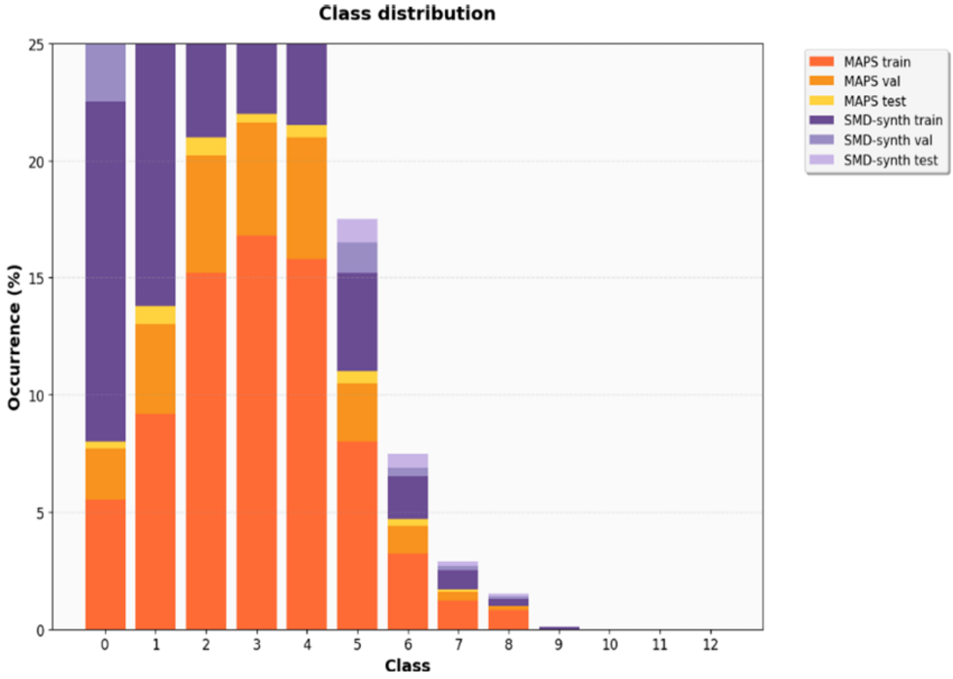
Figure 8 The Retina 1 dataset comprises the original image, ground-truth image, original matched filter, and segmented images for various parameter combinations (see online version for colours)



Notes: These combinations include s = 1, r = 6, s = 2, r = 10, s = 3, r = 18, s = 4, r = 28, and s = 5, r = 25. The technique used to segment the CDB dataset was a combination of the Fast Guided Filter and the Matched Filter.

We separated degrees into their own classes and determined their distributions using the polyphonic values from the datasets. Figure 9 shows the distributions in general. Frames with $d \geq 7$ make up fewer than 5% of each dataset, whereas frames with $d \geq 6$ nevertheless contribute to SMD-synth. Class 0 and 1 occurrences are substantially higher in SMD-synth compared to MAPS. There are two types of frames: those with no active pitch and those with a single pitch.

Figure 9 Division of Subsets in SMD-synth datasets and MIDI-aligned piano sounds (MAPS) (see online version for colours)



Notes: With 12 simultaneous notes, the highest possible polyphony degree is 12 – the frequency of classes greater than 5 decreases rapidly. Further details of the subsets are provided in Table 3.

This disparity prompted us to look at three distinct LPE class partition algorithms. We examined three categories in the first plan: monophony (one note), polyphony (two notes active), and silence (zero notes). Class 2 was thus defined as all degrees of polyphony $d \geq 2$. Detecting more nuanced degrees of polyphony was the goal of the second technique, which involved six classes. Regarding the difference for polyphonic degrees $d \geq 6$, we previously stated that all degrees $d \geq 5$ were categorised as class 5. Then, you may continue with classes 0 and 1, while classes 2–4 demonstrated different degrees of polyphony by playing 2–4 notes at once, accordingly. After that, class 5 merged all advanced degrees once more. The third tactic was to take into account all 13 degrees of polyphony, which were the end consequence of the maximum polyphony of 12 in the two sets of data. Our goal in implementing this third technique was to identify the areas with the most significant decline in forecast accuracy.

Table 3 Three-set division (training, validation, testing) of datasets from MAPS Configuration 2 and SMD-Synth

Dataset	*Number of files per set*	*Duration (train/val/test) [min]*	*Max polyphony (train/val/test)*
MAPS Config. 2	180/30/60 (270)	711.0/133.5/261.7	12/11/12 (@44100/511)
			12/11/12 (@22050/511)
SMD-synth	35/7/8 (50)	201.3/38.6/21.2	12/11/12 (@44100/511)
			12/11/11 (@22050/511)

5 Conclusions

This study demonstrates that conventional wisdom about piano instruction may be significantly altered with the application of AI and DL. By objectively evaluating multidimensional performance qualities, the suggested AI-driven piano accompaniment and teaching system effectively provides learners with tailored feedback. The accuracy of feature extraction is further enhanced by the hybrid matched and fast guided filter technique, which guarantees trustworthy performance measurement. The experimental results show that the model's performance is significantly improved across several musical dimensions when hierarchical perceptual information is incorporated. Additionally, the system is well-suited for distant music instruction due to the use of cloud-based infrastructures, which guarantee scalability and accessibility. Automatic creation still has a ways to go before it can ensure things like emotional depth and musical coherence, but the results suggest that piano lessons have made great strides. Improving auditory-led teaching approaches, developing more effective generative models for accompaniment, and expanding the system's applications to other instruments are all potential areas for future research. In conclusion, the study presented here offers a fresh and valuable approach to the problem of intelligent, adaptable, and time-saving piano instruction.

Declarations

All authors declare that they have no conflicts of interest.

References

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O. and Farhan, L. (2021) 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions', *J. Big Data*, Vol. 8, No. 1, p.53.

Dash, S. (2022) 'Guidance image-based enhanced matched filter with modified thresholding for blood vessel extraction', *Symmetry*, Vol. 14, No. 2, p.194.

Kale, A. and Altun, O. (2024) 'An efficient identity-preserving and fast-converging hybrid generative adversarial network inversion framework', *Eng. Appl. Artif. Intell.*, Vol. 138, Article 109287, DOI: 10.1016/j.engappai.2024.109287.

Kan, R.Y.P. (2022) *In-between Space of Learning: Professionalism in the Arts*, EdD Dissertation, National Institute of Education, Nanyang Technological University, Singapore.

Karamatlı, E. (2024) ‘Source separation and classification using generative adversarial networks and weak class supervision’, *Digit. Signal Process.*, Vol. 154, Article 104694, DOI: 10.1016/j.dsp.2024.104694.

Li, H. (2020) ‘Piano automatic computer composition by deep learning and blockchain technology’, *IEEE Access*, Vol. 8, No. 10, pp.188951–188958.

Li, W. (2022) ‘Analysis of piano performance characteristics by deep learning and artificial intelligence and its application in piano teaching’, *Front. Psychol.*, Vol. 12, No. 1, p.5962.

Liu, M. and Huang, J. (2021) ‘Piano playing teaching system based on artificial intelligence – design and research’, *J. Intell. Fuzzy Syst.*, Vol. 40, No. 2, pp.1–9.

Pang, S.E. (2024a) ‘Nuances of an in-between space of learning through auditory approaches in early piano instruction’, *Behav. Sci.*, Vol. 14, No. 12, p.1128.

Pang, S.E. (2024b) *Auditory Teaching Approaches in Early Piano Instruction: Case Studies in Singapore*, Bachelor’s thesis, Royal College of Music, London-Nanyang Academy of Fine Arts, Singapore.

Park, J. et al. (2024) ‘Piano performance evaluation dataset with multilevel perceptual features’, *Sci. Rep.*, Vol. 14, p.23002.

Peñalver Vilar, M. and Valles Grau, L. (2020) ‘Vocal piano accompaniment: a constant research towards emancipation (2)’, *English Lang. Lit. Cult.*, Vol. 5, No. 1, pp.25–35.

Phanichraksaphong, V. and Tsai, W-H. (2023) ‘Automatic assessment of piano performances using timbre and pitch features’, *Electronics*, Vol. 12, No. 8, p.1791.

Song, L. (2024) ‘Design and implementation of remote piano teaching based on attention-induced multi-head convolutional neural network optimized with hunter–prey optimization’, *Int. J. Comput. Intell. Syst.*, Vol. 17, No. 2, DOI: 10.1007/s44196-023-00379-3.

Stün, E. and Ozer, B. (2000) ‘Effects of piano accompaniment on instrument training habits and performance self-efficacy belief in flute education’, *Cypriot J. Educ. Sci.*, Vol. 15, No. 3, pp.412–422.

Yu, X. and Ma, N. (2023) ‘Developments and applications of artificial intelligence in music education’, *Technologies*, Vol. 11, No. 2, p.42.

Zhang, M. (2023) ‘Self-powered electronic skin for remote human-machine synchronisation’, *ACS Appl. Electron. Mater.*, Vol. 5, No. 1, pp.498–508.

Zheng, H. (2022) ‘Construction and optimisation of an artificial intelligence-assisted interactive college music performance teaching system’, *Sci. Program*, Vol. 2022, No. 1, p.3199860.

Zhou, D. (2025) ‘Generative adversarial network for adaptive piano accompaniment’, *Systems Softw. Comput.*, Vol. 2025, No. 2025, Article 200289.