Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **9**, Number **1**, February **2024**

Piano Monotone Signal Recognition based on Improved Endpoint Detection and Fuzzy Neural Network

Xing-Zhe Jiang*

Department of Humanities and Arts
Zhe Jiang business technology institute
Ningbo 315000, P. R. China
jiangxingzhe2023@163.com

Sha Na

Belt and Road Research Institute
Graduate University of Mongolia
Ulaanbaata 14200, Mongolia
565056470@163.com

*Corresponding author: Xing-Zhe Jiang
Received August 3, 2023, revised October 20, 2023, accepted December 27, 2023.

Abstract. *: Piano music recognition techniques involve knowledge theories related to computer multimedia, signal processing and artificial intelligence. Piano music recognition is of great practical value to the beginner pianist. Piano music recognition enables the identification of audio files generated from piano playing and comparison with standard electronic scores. In this paper, a novel endpoint detection method is combined with fuzzy neural networks for the recognition of piano playing music. Firstly, Mel-Frequency Cepstral Coefficients (MFCC) is selected as a feature of the monophonic signal and the audio files are analyzed in the time-frequency domain. An endpoint detection algorithm based on instantaneous power variation is proposed to address the drawback that the traditional double-threshold endpoint detection algorithm relies too much on the threshold value. By finding the peak of the instantaneous power variation in order to determine the starting point of the note, and designing a two-level judgment to determine the endpoint corresponding to each starting point. Then, to address the problem that fuzzy neural networks cannot extract data features in depth, this paper proposes an improved fuzzy neural network-based piano single note recognition method by effectively combining convolutional neural networks with fuzzy neural networks. The convolutional layer is used to extract the data features in depth, and the pooling layer is used to reduce the dimensionality of the extracted features. The experimental results show that the proposed method can achieve an accuracy of 97.82% for the recognition of 88 single tones of the piano. The work in this paper has good reference value for the application of artificial intelligence in automatic music recognition.*
Keywords: Piano monophonic recognition; Endpoint detection; Artificial intelligence; Fuzzy neural networks; Convolutional neural networks

1. **Introduction.** Piano music recognition technology involves knowledge of theories related to computer multimedia, signal processing and artificial intelligence. Piano music recognition is of great practical value to the beginner pianist. Piano music recognition is able to identify the audio files generated by piano playing and compare them with standard electronic scores. As living standards improve, people are paying more and

more attention to the development of musical literacy [1,2]. With the promotion of artificial intelligence technology and machine learning technology that have emerged in recent years, piano music recognition, as a branch of pattern recognition technology, has been developed rapidly [3,4].

When used as an audio sign, music can convey emotions and thoughts. It is also a mode of entertainment that can enrich people's spiritual life. With the development of electronic science and computer science, people's research on music has started to develop towards signal processing and pattern recognition. A brand-new cross-disciplinary field called computer automatic music recognition uses expertise from many other fields, including physics, signal processing, artificial intelligence, music theory, and many more, to conduct its study [5]. A global wave of artificial intelligence learning has been sparked by the quick growth of Internet technology and computers since the turn of the twenty-first century. The use of artificial intelligence technology has quickly impacted many fields of study and everyday life. For example, intelligent transportation technology [6], fingerprint recognition technology [7], license plate detection technology [8] and so on.

Speech recognition and music signal recognition have also gradually become research hotspots in the field of artificial intelligence [9], such as Siri from Apple in the US and various music apps. Speech recognition functions have all brought convenience to people's daily life. Music recognition applications also reflect a strong market demand. The piano is the most popular instrument for learning music because it is the monarch of all instruments. However, learning to play the piano requires a lot of practice and skilled instruction. The modern world's hectic pace makes it difficult for people to find the time for formal education, and the scarcity of qualified piano instructors generally drives up the cost [10].

More and more academics are devoting their lives to the study of music in order to address the aforementioned difficulties. Simulating music and music recognition have also grown in popularity as study areas for artificial intelligence. Automatic melody and music score recognition has also advanced and achieved some significant strides. Research on recognition of piano playing music originated from speech recognition technology, which started much earlier [11]. Most traditional music recognition methods are based on Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). For a long time, the method of building music recognition models based on GMM-HMM was the mainstream approach in the related field [12,13]. However, this traditional method also has problems such as difficulty in describing complex acoustic models and weak model classification ability [14]. To address these problems, some researchers have combined fuzzy inference systems with artificial neural networks to construct a fuzzy neural network model (FNN), and applied the model to the field of speech recognition.

Therefore, the purpose of this study is to combine the new endpoint detection technology with the improved fuzzy neural network model to accurately identify the pitch and duration of 88 notes in piano playing music. The study of music recognition based on fuzzy neural networks will help to improve the theory of fuzzy neural networks and provide new ideas and methods for music recognition techniques.

1.1. **Related Work.** In terms of technology, there are many similarities between speech recognition and music recognition. The purpose of speech recognition is to convert audio signals into corresponding language and text content, while music recognition is to convert them into corresponding notes and tunes [15]. When employing voice recognition, the initial step is to pick out individual words from continuous speech, and then, using context, integrate those words into a whole phrase. In order to recognize music, the initial phase also entails picking out individual notes from continuous sounds and putting them all

together to form a whole composition. For more difficult instances, music recognition also requires the recognition of chords, instruments and other information. In general, endpoint detection techniques are a fundamental requirement for both speech recognition and music recognition.

The purpose of endpoint detection technology is to accurately identify the starting point and ending point of each word/note from continuous audio, which is the basis of speech recognition and music recognition. Research on endpoint detection algorithms dates back to the 1950s. Endpoint detection algorithms can be divided into two main categories according to their principles [16]: the first category is based on the threshold value approach, which distinguishes the speech segment/music segment from the noise segment by analyzing the sound signal in the time-frequency domain and setting the feature parameter threshold value; the second category is based on the pattern recognition approach, which detects the corresponding features of the sound signal by constructing a model to match them.

An endpoint identification technique based on the average amplitude difference was proposed by Ross et al. [17] in 1974, and it has the advantages of being simple to use and minimal in complexity. However, the false detection rate is considerable when the sound signal's amplitude fluctuates often. The short-term likelihood function has an extreme value at integer multiples of the period and the same length as the source. Therefore, Rabiner and Herrmann [18] proposed a classical double-threshold endpoint detection algorithm based on instantaneous power. Since the instantaneous power of speech turbulence is larger than that of noise, the double-threshold endpoint detection algorithm has been widely used with good accuracy in high SNR environments. To address the problem of poor detection under low SNR (Signal Noise Ratio), An endpoint identification technique based on instantaneous synchronization and zero-crossing rate was proposed by Li et al. [19], and it likewise obtained good reliability in a low SNR.

Artificial intelligence has recently been used to improve voice recognition and music identification. An end-to-end voice recognition model built on an attention mechanism was proposed by Watanabe et al. [20]. In order to address the issue of low voice recognition rate at low SNR, Tian [21] presented a voice recognition approach based on convolutional neural networks. A voice recognition technique suggested by Palaz and Magimai-Doss [22] makes speech recognition more noise-resistant by extracting signal properties using deep neural networks. In the field of music recognition, Borde et al. [23] chose Mel-Frequency Cepstral Coefficients (MFCC) as the feature parameter and used BP neural network as the matching model to achieve high recognition accuracy. Liu et al. [24] proposed an audio recognition method based on extreme learning machine.

1.2. **Motivation and contribution.** To sum up, both speech recognition and music recognition have made great progress in the decades of development. In endpoint detection, the traditional algorithm theory is relatively mature, but there are still some shortcomings, such as over-reliance on threshold and inability to find a balance between accuracy and complexity. Meanwhile, fuzzy neural networks, which are made by combining artificial neural networks with fuzzy inference systems, have excellent mass data training capability and model classification ability [25,26].

Therefore, this paper takes the audio files generated from piano playing music as the research object, and focuses on the pitch and time value recognition technology of notes. To address the shortcomings of traditional double-threshold endpoint detection algorithms [27], which are overly dependent on thresholds and have poor noise immunity, this research suggests a instantaneous power variation-based endpoint detection technique. In the pitch recognition part, this paper achieves a high accuracy rate by drawing the MFCC feature

map of each note and inputting it into an improved fuzzy neural network for classification, so as to achieve the recognition of note pitches. The main innovations and contributions of this paper include.

(1) To address the drawback that the traditional double threshold endpoint detection technology relies too much on the threshold value, an endpoint detection algorithm based on the instantaneous power variation is proposed. The dependency on the threshold is decreased by locating the peak value of the instantaneous power change to establish the starting point of the note and constructing a two-stage judgment to establish the end point corresponding to each starting point.

(2) To address the problem that fuzzy neural networks have difficulty in extracting data features in depth, an improved fuzzy neural network model based on convolutional neural networks is proposed. First, the data features are extracted in depth by using a convolutional layer. Then, the pooling layer is used to reduce the dimensionality of the extracted features. Finally, the data output from the pooling layer is fuzzified. The fuzzy inference system calculates the applicability of the current rule based on the subordination value of the input signal that has been fuzzed. After completing the data defuzzification, the output layer will output the music recognition results.

2. **Basic audio signal analysis.**

2.1. **Basic characteristics of music.**

The essence of sound is the sound waves produced by the vibration of objects. It is transmitted through various media and is perceived by the auditory organs of humans and other animals. Music is a form of art in which various types of sound are combined to convey people's thoughts and feelings. With the development of human society as a whole, music has developed into a multi-disciplinary system. The system of musical characteristics consists of three types: overall characteristics, basic characteristics and complex characteristics. Complex characteristics are expressed in rhythm, melody, harmony, etc. The overall characteristics are expressed in musical style, emotional connotation, etc. Understanding the basic characteristics of music is the first step in music recognition. The basic characteristics of music are expressed in four properties: (1) Pitch is the frequency of vibration of a sounding body per unit of time. (2) Duration is the duration of the vibration of a sound body. (3) Volume is the amplitude of the vibration of the vocal body. (4) Timbre is the distribution of overtones when the vocal body vibrates.

The four properties mentioned above play different roles in musical expression. Pitch and duration are the two most prominent of these properties. Melody is mainly composed of two elements: sound length and pitch. For the same piece of music, using different timbre or volume, its basic melody characteristics are unchanged. However, if the pitch and duration are slightly modified, the whole music will be destroyed. Pitch and duration are of great importance in music recognition and are a major challenge in music recognition. Pitch involves calculating the fundamental tone of each note. The duration can be obtained by calculating the duration of each note. The fundamental tone is also known as the dominant frequency, which has a physical meaning as the eigenfrequency. The dominant frequency is the lowest frequency of the sound waves formed by the vibration of an object. A note has an overtone in addition to the fundamental tone. Overtones are also known as harmonic frequencies, the physical meaning of which is the resonant frequency. The harmonic frequency is usually an integer multiple of the fundamental frequency. The different distributions of overtones make up the different timbres, but the fundamental tone is the same for all timbres. The main task of pitch identification is therefore to find the fundamental tone from the many overtones.

The piano is a keyboard instrument. Keyboard instruments, in general, use twelve mean meters, so that the ratio of the frequencies of two adjacent keys is 21/12. The frequency relationship of the notes of a piano is shown in Figure 1.

$$f_2/f_1 = 2^{1/12} \tag{1}$$

The piano is made up of 52 white keys and 36 black keys, a total of 88 keys. The piano

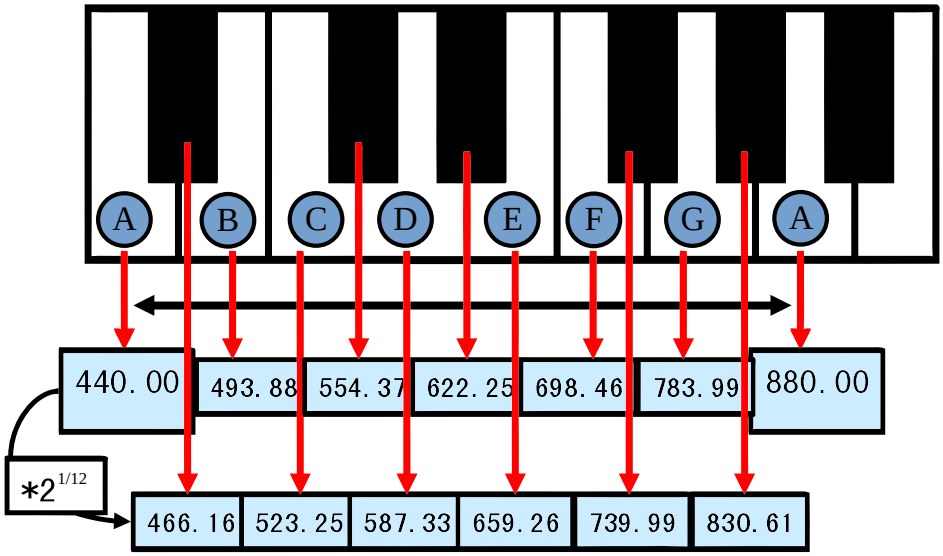

Figure 1. Note frequency relationships in the piano

covers a range of fundamental frequencies from 27.5Hz to 4186Hz, which includes almost the entire range of pitches in the musical system. The free vibration response of the piano is shown as follow:

$$y(x,t) = \sum_{i=1}^{n} \left\{ \frac{2}{l} \left[\int_0^l f_1(x) \sin\left(\frac{n\pi x}{l}\right) dx \right] \cos(p_n t) + \frac{2}{nc\pi} \left[\int_0^l f_2(x) \sin\left(\frac{n\pi x}{l}\right) dx \right] \sin(p_n t) \right\} \sin\left(\frac{p_n x}{c}\right) \tag{2}$$

The natural frequency of the system is shown as follow:

$$p_n = \frac{n\pi}{l} \sqrt{\frac{T_0}{\rho A}} (n = 1, 2, 3 \ldots) \tag{3}$$

where T_0 represents the tension of the chord, A represents the cross-sectional area of the chord, ρ represents the density of the chord, l represents the length of the chord, and n represents the total number of intervals.

2.2. **Windowing and framing of monophonic signals.** This article uses a computer to process the music. The input music is captured using a microphone or recorder. This process is the sampling and quantization of the music signal. A music signal can be thought of as an analogue signal whose amplitude varies continuously over time and which must be sampled and quantized. We need to convert analog signal of music into a digital signal. Generally speaking, the higher the sampling frequency, the lower the distortion of the original signal. For music signal processing, the sampling frequency is usually 8-10 kHz. The highest fundamental frequency of the piano music studied in this paper is 4186Hz, so all the sampling frequencies involved in this paper, Fs, are 48000Hz, which satisfies the sampling theorem.

The digitized music signal is in fact a time-varying signal. As the instrument is in a constant state of change during the sounding process, the music signal in practice can be

seen as a linear time-varying signal. However, this variation is very slow for the frequency of the signal, so it can be assumed that the signal is smooth for a very short period of time. Usually this time period is taken to be between 10 and 30 ms, and all subsequent analyses in this paper have been carried out under this assumption.

The frame-splitting plus windowing operation uses a window function that slides over the sampled signal and splits it into a number of instantaneous signals [28]. These instantaneous signals can be considered as smooth signals. Each instantaneous signal is a frame, and its length is the frame length. The length of each frame is generally taken to be 10-30 ms. There are various options for the window function and different window functions can affect the final analysis results. At present, there are three main window functions in common use: rectangular, Hanning and Hamming windows. In the time-frequency domain analysis of music signals, the choice of window function is very important. Although the rectangular window has excellent frequency recognition accuracy and spectral smoothness, the amplitude recognition accuracy is low and the waveform details are easily lost. Leakage can occur with the Hanning window. In contrast, the Hamming window is the most widely used window function, as it can overcome leakage and has a high accuracy in the time-frequency domain.

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], 0 \le n \le N-1 \\ 0, \text{others} \end{cases} \tag{4}$$

where N is the window length.

The length of the window function plays a crucial role in accurately reflecting the amplitude changes of the signal. If the window length N is too large (more than one pitch period), then the window function is equivalent to a very narrow low-pass filter, and therefore can not fully reflect the details of the waveform changes. Conversely, if the window length N is too small (less than one pitch period), then the window function is equivalent to a widened filter, and therefore does not provide smooth information about the waveform changes. It is generally accepted that a frame should contain between 1 and 7 fundamental periods. Therefore, for the 48000 Hz sampling frequency used in this paper, a frame length of 10 to 30 ms was chosen.

3. **Improved endpoint detection algorithm on instantaneous power variation.**

3.1. **Time-frequency feature extraction of signals.**

At present, the traditional audio signal analysis methods are time domain analysis and frequency domain analysis. The waveform of time domain analysis is susceptible to noise interference. On the contrary, frequency domain analysis is robust to noise interference and can better represent the acoustic characteristics of audio signals. This is why most practical music signal analysis uses frequency domain analysis techniques. Frequency domain analysis is a technique that transforms the time domain signal into the frequency domain and extracts the cepstrum coefficients from the music signal. The cepstrum coefficients are used to estimate the fundamental period.

The short time Fourier analysis of a music signal can be used to obtain a sound spectrum diagram of the signal. A sound spectrum is a three-dimensional spectrum with time on the horizontal axis and frequency on the vertical axis. The spectrogram combines the characteristics of a frequency spectrum and a time domain waveform, and clearly shows the change of the sound signal spectrum over time. The process of generating a sound spectrum diagram is shown in Figure 2.

After converting the physical frequencies to Mel frequencies according to certain rules, we can extract the coefficients used to characterize the musical properties in the Mel frequency domain, which are the MFCC (Mel-Frequency Cepstral Coefficients) [29]. In

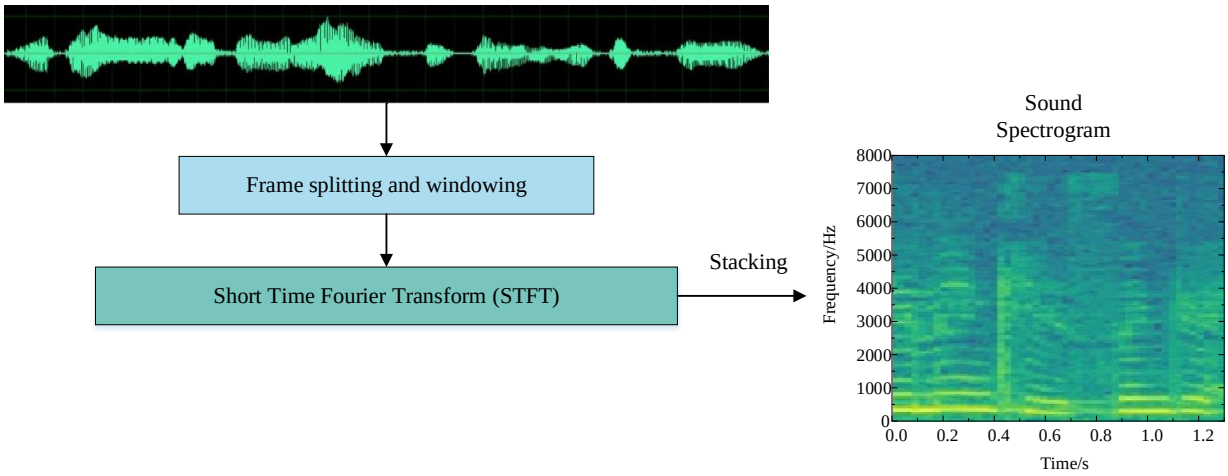

Figure 2. Extraction process for acoustic spectrograms

contrast to LPCC (Linear Predic-tion Cepstrum Coefficient), which is based on the theory of human ear hearing, MFCC does not correlate with the nature of the input signal. MFCC is able to avoid errors arising from differences between the approximation and the actual value. The core of the work in this paper is the recognition of note pitches using a neural network model, and therefore MFCC, which conforms to the properties of human ear hearing, has an advantage over acoustic spectrograms and LPCC.

Firstly, in order to extract the MFCC, the input speech signal needs to be pre-emphasized by passing it through a high-pass filter. In this paper, the pre-emphasis factor is set to 0.97. Secondly, the monophonic signal is windowed and framed according to the method described above. For the framed and windowed music signal, a Fast Fourier Transform (FFT) is also performed to obtain the energy distribution on the spectrum.

$$X_{\mathrm{a}}(k) = \sum_{\mathrm{n}=0}^{N-1} x(n) e^{-j2\pi k/N} \tag{5}$$

where *x(n)* is the input music signal, N is the number of points in the Fourier transform and *X(k)* is the representation of the signal in the frequency domain.

In order to obtain a smooth spectrum and reduce the effect of harmonics generated, the energy spectrum needs to be filtered using a triangular bandpass filter. In this paper a set of triangular band-pass filters with M filters is used. The number of filters, M, is usually taken to be between 22 and 26.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{2(\mathrm{k}-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m) \le k \le f(m+1) \\ 0 & k \ge f(m+1) \end{cases} \tag{6}$$

where $H_m(k)$ is the frequency response of the triangular filter and *f(m)* is the center frequency. The logarithmic energy *S(m)* is then calculated after filtering.

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right) \tag{7}$$

The MFCC is obtained by performing a discrete cosine transform (DCT) on *S(m)*.

$$\mathrm{C}(n) = \sum_{m=0}^{N-1} S(m) \cos\left(\frac{(n\pi(m-0.5)}{M}\right) \tag{8}$$

The flow of extracting MFCC parameters is shown in Figure 3.

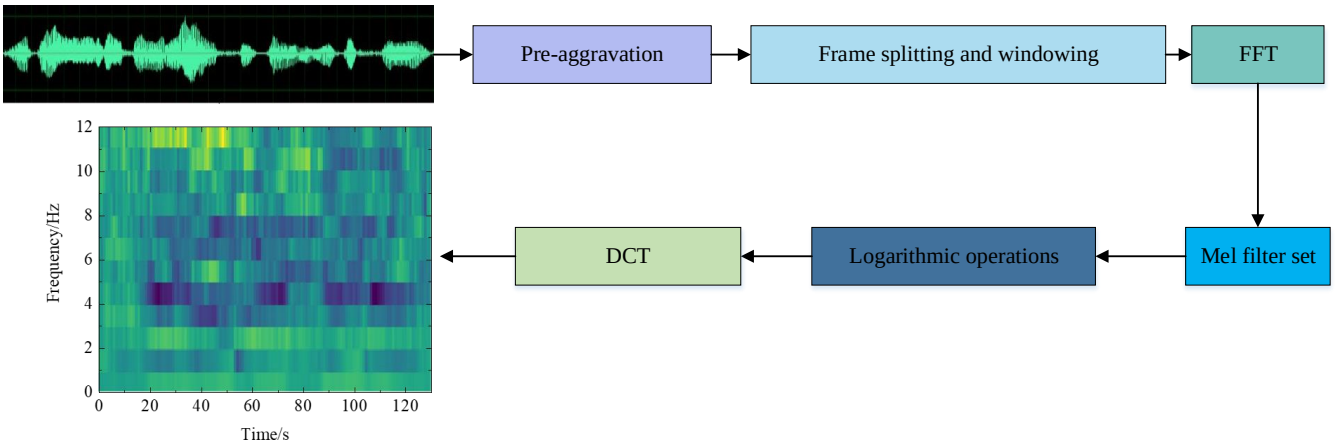

Figure 3. Flow of extracting MFCC parameters

3.2. **Endpoint detection on instantaneous power variation.** Endpoint detection extracts the beginning and ending points of a music signal from a piece of audio, and thus obtains the note duration characteristics of each note. Endpoint detection is one of the most important steps in the field of music recognition.

The double threshold method is a commonly used endpoint detection algorithm based on the signal time domain feature parameters [30]. The first level of judgment in the double threshold method uses mainly instantaneous power feature parameters. First, a higher threshold M_1 is set based on the overall instantaneous power profile of the audio signal. From there, a lower threshold M_2 is set based on the average instantaneous power value of the ambient background audio signal. The average over-zero rate of the first few frames of the audio signal is taken as the threshold M_3.

Based on the above steps it can be found that the double threshold detection algorithm relies heavily on the setting of three thresholds. The accuracy of the algorithm is directly related to the good or bad threshold settings. To address this drawback, this paper proposes to improve the double threshold detection algorithm by using instantaneous power variation. The main idea is to find the information of energy mutation through the characteristic of instantaneous power change, so as to determine the starting point of notes. Then, according to each starting point, two levels of judgment are designed to finally determine the corresponding end point of each starting point.

Firstly, each note of the piano can be seen as an instantaneous power pulse. The instantaneous power of the input voice information is computed after it has been windowed and framed.

$$E_i = \sum_{n=0}^{L-1} |x(n)|^2 \tag{9}$$

where *x(n)* denotes the values of the n-th position in the i-th frame of the signal and L denotes the window period.

The size of the window period is related to the sampling frequency. In this paper the sampling frequency is 48000 Hz and the window length is 2% of the sampling frequency, i.e. 960 sampling points.

Then, the instantaneous power variation among two neighbouring frames is calculated.

$$\Delta E_i = E_i - E_{i-1} \tag{10}$$

The instantaneous power change calculates the energy difference between frames, not the energy difference between every two sampling points. This method therefore has a smoothing effect on the calculation of the overall energy of the audio signal.

To verify the advantages of the proposed algorithm over the traditional double threshold algorithm, 10 sets of piano music were recorded for comparison, with a total of 792 note samples. Both the traditional double threshold endpoint detection algorithm and the instantaneous power variation based endpoint detection algorithm were used to detect the endpoints of the two sets of recordings, and the results are shown in Table 1. It can be seen that the improved endpoint detection algorithm proposed in this paper is more resistant to noise interference and the accuracy of note recognition is higher.

Table 1. Comparison of endpoint detection results

	Actual number of notes	**Correct detection of the number of notes**	**Accuracy**
Traditional Double Threshold Algorithm	792	720	90.9%
Instantaneous power variation algorithm	792	787	99.4%

4. Single-tone signal recognition based on improved fuzzy neural networks.

4.1. Fuzzy reasoning systems.

The traditional acoustic model is the GMM acoustic model, which is a shallow model with poor recognition performance in the case of insufficient audio signal samples, and this has seriously hindered the development of audio recognition technology. Therefore, this paper tries to combine convolutional neural networks with fuzzy neural networks effectively to solve some problems of traditional recognition methods.

Each layer of the network of a fuzzy neural network corresponds to a sub-module in the fuzzy inference system [31]. In general, the fuzzy inference system of FNN consists of four parts: fuzzification, knowledge base, logical judgment and defuzzification. The structure of the fuzzy inference system is shown in Figure 4.

4.2. Fuzzy neural networks.

Neurons of FNN can be divided into two types [32]: fuzzy neurons with clear inputs and fuzzy neurons with fuzzy inputs. Elements. Neurons with clear inputs use an affiliation function to complete weighting of the input data. Fuzzy neurons with fuzzy inputs do not use the affiliation function to process the data, but directly correct the input data. In this work, fuzzy neurons with clear inputs are chosen to build a piano monophonic signal recognition system. The structure of the fuzzy neuron with clear input is shown in Figure 5. The fuzzy neurons are calculated as shown as follow.

$$y\left(x_1, x_2, x_3, \ldots, x_n\right) = \mu_1\left(x_1\right) \otimes \mu_2\left(x_2\right) \otimes \ldots \otimes \mu_n\left(x_n\right) \tag{11}$$

where $\otimes$ denotes the cumulative operator, x_n denotes the n-th input to the neuron, y denotes the output, and $\mu_n(x_n)$ denotes the affiliation function of the n-th weight.

The fuzzy neural network constructed in this paper has a total of five layers. The first layer is the input layer. The second layer is the fuzzification layer. The third layer is

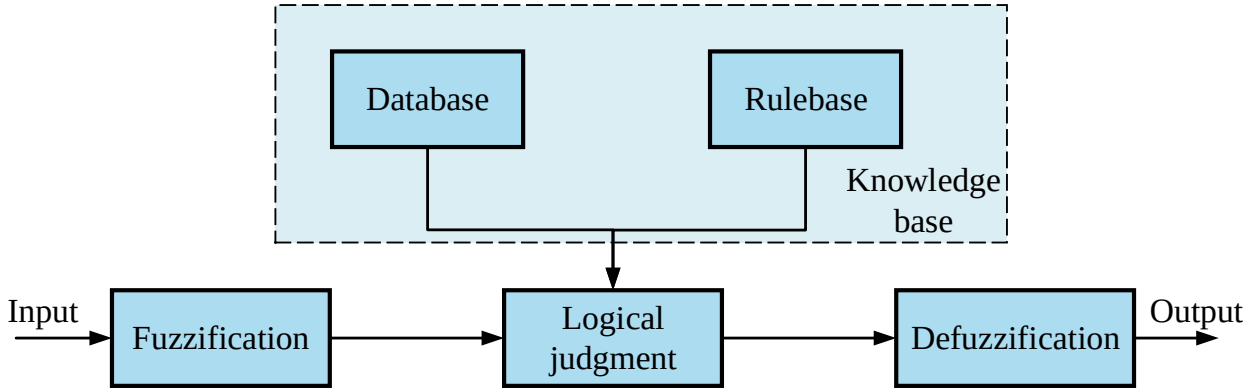

Figure 4. Fuzzy inference system architecture

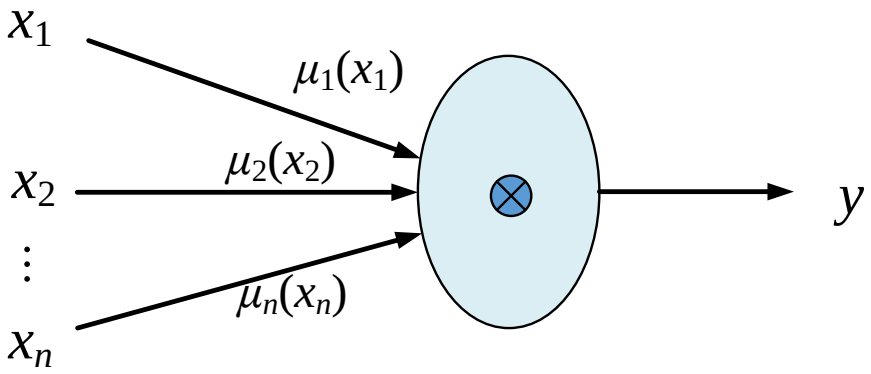

Figure 5. Structure of a fuzzy neuron with clear inputs

the fuzzy inference layer. The fourth layer is the defuzzification layer. The fifth layer is the output layer. The affiliation degree of each input feature component is calculated according to the fuzzy rules.

$$\mu(x) = e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{12}$$

The resulting output values are normalized and the error e is calculated.

$$Y_i = \frac{\sum_{i=1}^{n} y^i \left(p_0^i + p_1^i x_1 + \cdots + p_i^i x_t\right)}{\sum_{i=1}^{n} y^i} \tag{13}$$

$$e = \frac{1}{2}(Y - Y_d)^2 \tag{14}$$

In addition, the structure of fuzzy neurons needs to be updated.

$$p_j^i(t) = p_j^i(t-1) - \alpha \frac{\partial e}{\partial p_j^i} \tag{15}$$

$$\frac{\partial e}{\partial p_j^i} = (Y_d - Y_i)\, y^i / \sum_{i=1}^{n} y^i x_j \tag{16}$$

$$\mu_j^i(t) = \mu_j^i(t-1) - \beta \frac{\partial e}{\partial \mu_j^i} \tag{17}$$

$$\sigma_j^i(t) = \sigma_j^i(t-1) - \beta \frac{\partial e}{\partial \sigma_j^i} \tag{18}$$

4.3. **Convolutional-Fuzzy Neural Networks.** Convolutional neural network (CNN) is a deep neural network that mimics the mechanism of object perception by human brain and has good performance in speech recognition, image processing, etc. The most core hidden layers of CNN have two layers [33]: convolutional layer and pooling layer. the structure of CNN is shown in Figure 6. The convolutional and pooling layers are the core parts of CNNs to improve the power of feature data representation and reduce the complexity of models. Considering the advantages and disadvantages of the CNN model

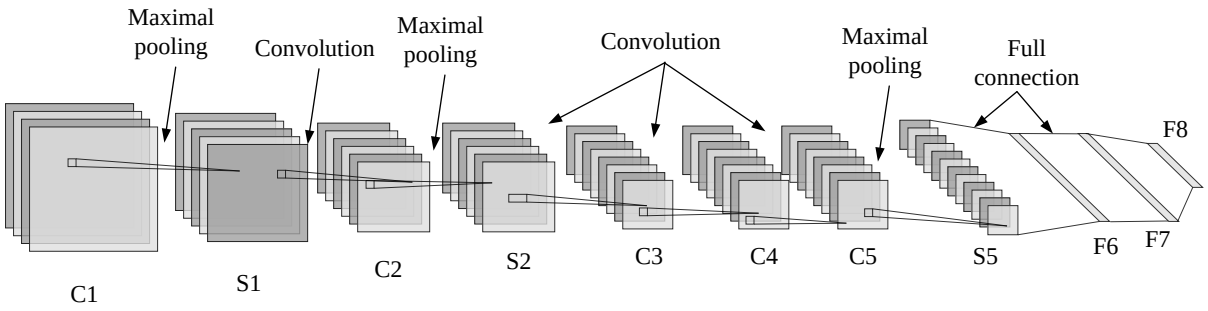

Figure 6. CNN structure

and the FNN model, this paper proposes to introduce the convolutional and pooling layers into the FNN to construct a Convolutional-Fuzzy Neural Network (C-FNNN) based model for monophonic signal recognition. The model can make use of the CNN model to improve the representational power and reduce the dimensionality capability of the data, while having the fuzzy data processing capability of the FNN model. In this paper, a seven-layer C-FNN is constructed, and the specific structure is shown in Table 2. When

Table 2. Structure of C-FNN

N-th floor	Type	Number of neurons	Size of the nucleus
1	Input layer	1	-
2	Convolutional layer	64	8 x 8
3	Maximum pooling layer	64	2 x 2
4	Fuzzification layer	64	-
5	Fuzzy reasoning layer	34	-
6	Defuzzification layer	34	-
7	Output layer	1	-

the input piano audio signal has been pre-processed and MFCC parameters extracted, it is fed into the convolution layer. After completing the depth extraction of the feature data, the data is input to the maximum pooling layer. After completing the maximum pooling calculation, the data is input to the fuzzification layer. In the fuzzy inference layer, the applicability of the current rule is deduced by matching judgment on the affiliation degree. When the data is transferred to the deblurring layer, the normalization calculation is completed by softmax function. Finally, the output layer outputs the single tone signal recognition result.

The connection between two neurons in adjacent hidden layers is called a channel.

$$v_{ljj}(K, Y) = \sum_k \mathrm{vec}\left(\mathbf{K}_{kl}\right) * \mathrm{vec}\left(\mathbf{Y}_{kij}\right) \tag{19}$$

where $\mathbf{K}_{kl}$ is the core matrix corresponding to the input channel k and the output channel l, $\mathbf{Y}_{kij}$ is the core matrix of the input samples (i, j) in the input channel k, and $\mathrm{vec}(\cdot)$ is the vector with the columns in the matrix arranged in sequence.

The data of convolution calculation will be input to the pooling layer.

$$v_{lij}(X) \leftarrow \max\left(X_{lij}\right) \tag{20}$$

$$v_{lij}(X) \leftarrow \frac{1}{K_r \times K_c} \sum_{m,n} x_{l,i_m,j_n} \tag{21}$$

The loss function of single tone signal recognition adopts cross entropy. Cross entropy describes the distance between two probability distributions, and the smaller the cross entropy is, the closer the two probability distributions are.

$$H(p, q) = -\sum_{x} p(x) \log q(x) \tag{22}$$

where p is the correct result and q is the predicted value.

The training flow of C-FNN single-tone signal recognition is shown in Figure 7. The testing flow of C-FNN single-tone signal recognition is shown in Figure 8.

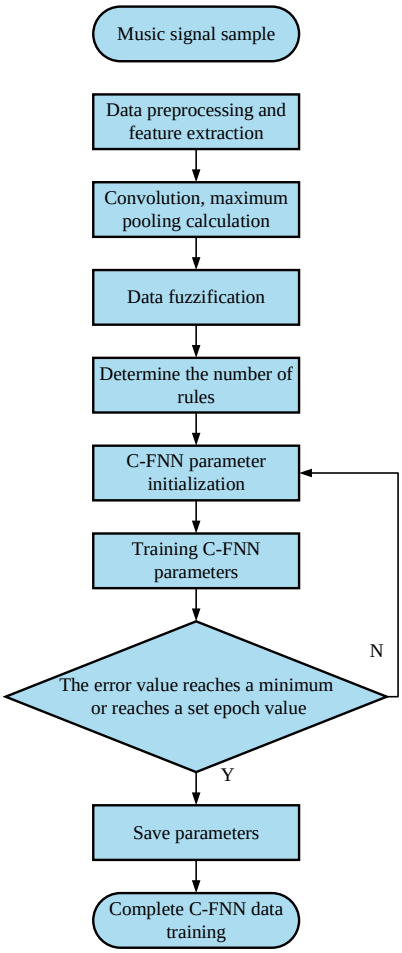

Figure 7. Training flow

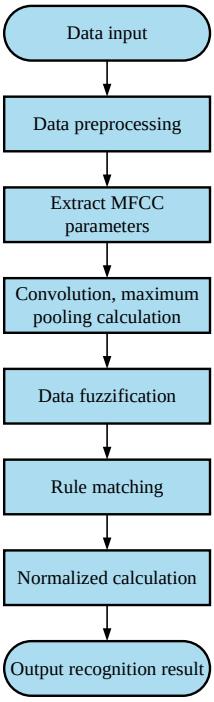

Figure 8. Testing flow

5. **Experimental results and analysis.**

5.1. **Hardware and software environment for the experiments.** In this paper, 88 single notes of data collected from a normal piano were used. The sample sampling rate was 48000 Hz. Piano monophonic recognition based on MFCC feature maps and C-FNN were implemented on the Windows 10 platform using MALAB, as shown in Table 3. The sampling time is five seconds. The file format is MIDI.

Table 3. Hardware environment

Configuration	**Parameters**
Processor	Intel Core I9-9900K
Memory	32GDDR4
Hard Disk	Kingston 240G Solid State Drive
GPU	NVIDIAGTX3060

5.2. **Comparison results with other deep neural network models.** The learning rate parameter determines whether the objective function can converge to a local optimum and the speed of convergence, and is one of the key parameters affecting the performance of the system. Therefore, choosing a suitable learning rate can effectively improve the comprehensive performance of single-tone recognition. First, to verify the advantages of using C-FNN, 10 sets of recordings (total 170 note samples) were subjected to single-tone recognition using CNN [34], FNN [35] and C-FNN methods respectively. All neural network models were trained 30 times and the results are shown in Figure 9.

When the learning rate was 0.1, all three deep neural networks did not reach the optimal value due to the excessive update of the weights, resulting in poor recognition. By gradually decreasing the learning rate, the recognition rate of all three deep neural networks improved. At a learning rate of 0.001, the recognition rate of C-FNN was 97.82%.

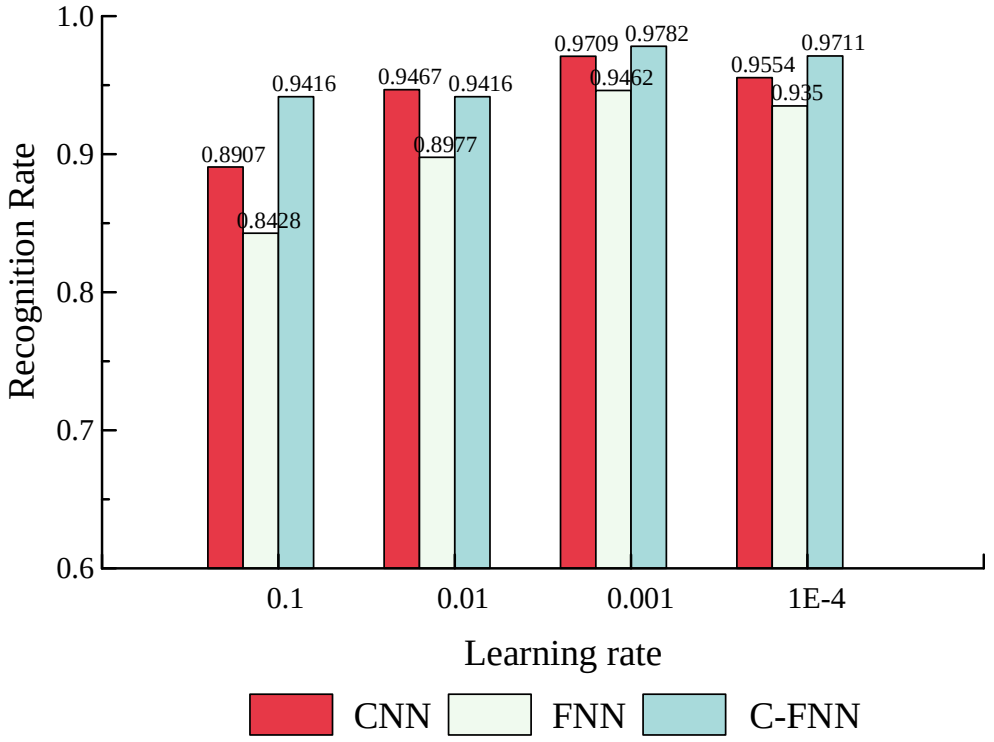

Figure 9. Effect of learning rate on recognition rate

Therefore, at a learning rate of 0.001, the best recognition results were achieved by the C-FNN-based single-tone recognition method. When the learning rate was reduced to 0.0001, the time spent on training the C-FNN samples increased sharply. On balance, the neural network model with a learning rate of 0.001 gives the best results.

When the learning rate is 0.001, the time taken by the three deep neural networks to complete a sample training is shown in Figure 10. The average time taken by CNN to complete a sample training is 75 seconds, which is the longest time. The average time for C-FNN to complete a data training is lower than that of CNN, with an average of 70 seconds. The average time for FNN to complete a sample data training is 5 seconds, and the sample training speed is much faster than the other two neural networks. In summary, the C-FNN-based single-tone recognition method is more comprehensive than the other two methods, proving its feasibility.

5.3. **Comparison results with conventional time-frequency domain methods.** Firstly, to verify the advantages of using C-FNN to identify pitches compared to the traditional time-frequency domain method, a piano recording containing four notes was recorded. The results of the three methods are shown in Table 4. The traditional time domain method is based on short-time amplitude difference function, and the traditional frequency domain method is based on wavelet analysis. It can be seen that the traditional

Table 4. Relative error comparison results

Pitch	Actual frequency (Hz)	C-FNN (%)	Traditional time domain method (%)	Traditional frequency domain method (%)
G4	392	2.069	4.961	3.603
D5	587.3	3.215	5.912	4.538
E4	329.6	2.423	100.001	5.011
C4	261.6	3.751	4.026	9.639

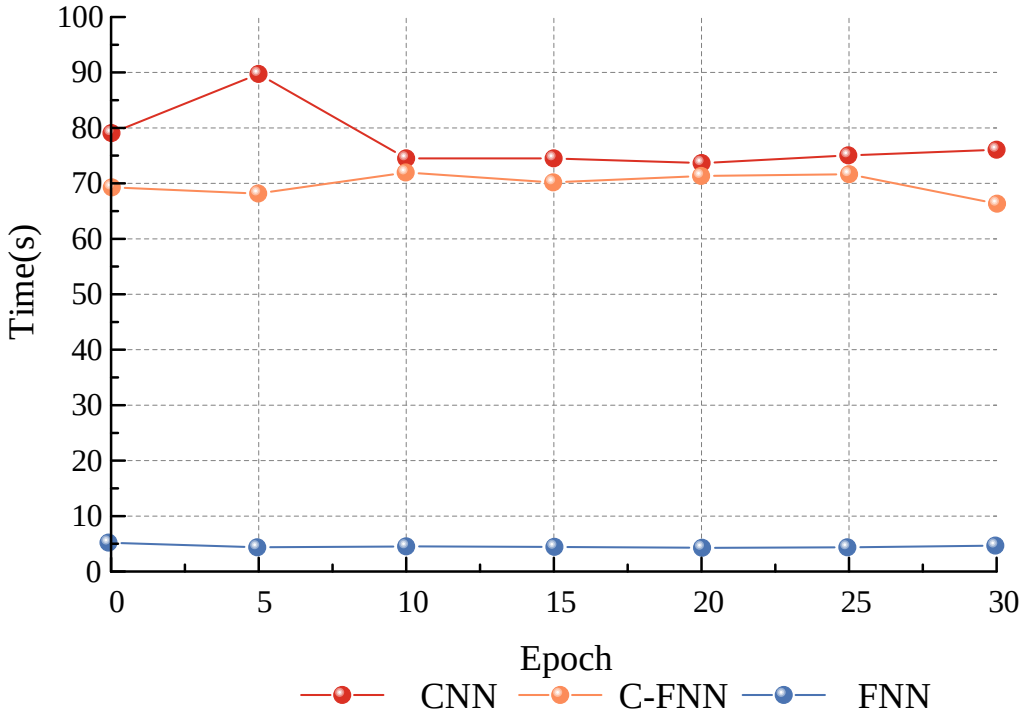

Figure 10. Training time for sample data

frequency domain method is able to detect 4 notes accurately. However, for the note C4, there is a large error of 9.639% in the calculation result of the traditional frequency domain method. The C-FNN algorithm accurately detects all four notes.

Then, 10 sets of recordings were identified using the three methods, and the final accuracy results are shown in Table 5. The experimental results show that the C-FNN algorithm proposed in this paper has the advantage of high recognition accuracy compared to traditional methods of time-frequency domain analysis.

Table 5. Pitch recognition accuracy

	C-FNN	**Traditional time domain methods**	**Traditional frequency domain methods**
Correct identification of the number of	157	113	138
Accuracy	97.82%	66.5%	81.2%

6. **Conclusion.** Piano music recognition technology involves knowledge of theories related to computer multimedia, signal processing and artificial intelligence. For piano beginners, piano music recognition is of great practical value. Therefore, this paper proposes a novel endpoint detection method with an improved fuzzy neural network model for piano monophonic signal recognition. First, an endpoint detection algorithm based on instantaneous power variations is proposed. By finding the peak of the instantaneous power variation to decide where the note should begin, and a two-level judgment is designed to determine each starting point's matching endpoint. Secondly, an improved fuzzy neural network model based on convolutional neural networks is proposed. The data features are extracted in depth by using a convolutional layer and the extracted features are dimensionalized by a pooling layer. After completing the data defuzzification, the output

layer will output the music recognition results. The experimental results show that the C-FNN achieves a 97.82% accuracy rate for single-tone recognition, which is higher than that of the traditional time-frequency domain analysis. The work in this paper has good reference value for the application of artificial intelligence in automatic music recognition.

REFERENCES

[1] J. Calvo-Zaragoza, J. H. Jr, and A. Pacha, "Understanding Optical Music Recognition," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1-35, 2020.

[2] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121-2133, 2022.

[3] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, 2339, 2023.

[4] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, 1977, 2023.

[5] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, "Hybrid hidden Markov models and artificial neural networks for handwritten music recognition in mensural notation," *Pattern Analysis and Applications*, vol. 22, no. 4, pp. 1573-1584, 2019.

[6] F. Yan, "Music Recognition Algorithm based on T-S Cognitive Neural Network," *Translational Neuroscience*, vol. 10, pp. 135-140, 2019.

[7] X. Chen, X. Qu, Y. Qian, and Y. Zhang, "Music Recognition Using Blockchain Technology and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 7025338-7025338, 2022.

[8] F. C. Thiesen, R. Kopiez, D. Müllensiefen, C. Reuter, and I. Czedik-Eysenberg, "Duration, song section, entropy: Suggestions for a model of rapid music recognition processes," *Journal of New Music Research*, vol. 49, no. 4, pp. 334-348, 2020.

[9] E. K. Wang, C.-M. Chen, M. M. Hassan, and A. Almogren, "A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain," *Future Generation Computer Systems*, vol. 108, pp. 135-144, 2020.

[10] K.-K. Tseng, J. Lin, C.-M. Chen, and M. M. Hassan, "A fast instance segmentation with one-stage multi-task deep neural network for autonomous driving," *Computers & Electrical Engineering*, vol. 93, 107194, 2021.

[11] Z. Xiao, X. Chen, and L. Zhou, "Real-Time Optical Music Recognition System for Dulcimer Musical Robot," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 23, no. 4, pp. 782-790, 2019.

[12] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, no. 4, pp. 365-389, 2017.

[13] A. Paul, R. Pramanik, S. Malakar, and R. Sarkar, "An ensemble of deep transfer learning models for handwritten music symbol recognition," *Neural Computing and Applications*, vol. 34, no. 13, pp. 10409-10427, 2021.

[14] W. Rodgers, F. Yeung, C. Odindo, and W. Y. Degbey, "Artificial intelligence-driven music biometrics influencing customers' retail buying behavior," *Journal of Business Research*, vol. 126, pp. 401-414, 2021.

[15] G. L. Wagener, M. Berning, A. P. Costa, G. Steffgen, and A. Melzer, "Effects of Emotional Music on Facial Emotion Recognition in Children with Autism Spectrum Disorder (ASD)," *Journal of Autism and Developmental Disorders*, vol. 51, no. 9, pp. 3256-3265, 2020.

[16] H. Tang, Y. Zhang, and Q. Zhang, "The Use of Deep Learning-Based Intelligent Music Signal Identification and Generation Technology in National Music Teaching," *Frontiers in Psychology*, vol. 13, pp. 762402-762402, 2022.

[17] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353-362, 1974.

[18] L. Rabiner, and O. Herrmann, "The predictability of certain optimum finite-impulse-response digital filters," *IEEE Transactions on Circuit Theory*, vol. 20, no. 4, pp. 401-408, 1973.

[19] J. Li, P. Zhou, X. Jing, and Z. Du, "Speech Endpoint Detection Method Based on TEO in Noisy Environment," *Procedia Engineering*, vol. 29, pp. 2655-2660, 2012.

[20] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240-1253, 2017.

[21] Y. Tian, "Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm," *IEEE Access*, vol. 8, pp. 125731-125744, 2020.

[22] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15-32, 2019.

[23] P. Borde, A. Varpe, R. Manza, and P. Yannawar, "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 167-175, 2014.

[24] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271-280, 2018.

[25] N. M. Ranjan, and R. S. Prasad, "LFNN: Lion fuzzy neural network-based evolutionary model for text classification using context and sense based features," *Applied Soft Computing*, vol. 71, pp. 994-1008, 2018.

[26] S. Hou, J. Fei, C. Chen, and Y. Chu, "Finite-Time Adaptive Fuzzy-Neural-Network Control of Active Power Filter," *IEEE Transactions on Power Electronics*, vol. 34, no. 10, pp. 10298-10313, 2019.

[27] H. Nasiri, and M. M. Ebadzadeh, "MFRFNN: Multi-Functional Recurrent Fuzzy Neural Network for Chaotic Time Series Prediction," *Neurocomputing*, vol. 507, pp. 292-310, 2022.

[28] J. Fei, and Z. Feng, "Fractional-Order Finite-Time Super-Twisting Sliding Mode Control of Micro Gyroscope Based on Double-Loop Fuzzy Neural Network," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 12, pp. 7692-7706, 2021.

[29] A. I. Siam, A. A. Elazm, N. A. El-Bahnasawy, G. M. El Banby, and F. E. Abd El-Samie, "PPG-based human identification using Mel-frequency cepstral coefficients and neural networks," *Multimedia Tools and Applications*, vol. 80, pp. 26001–26019, 2021.

[30] N. Yang, N. Dey, R. S. Sherratt, and F. Shi, "Recognize basic emotional states in speech by machine learning techniques using mel-frequency cepstral coefficient features," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 2, pp. 1925-1936, 2020.

[31] S. N. Qasem, and A. Mohammadzadeh, "A deep learned type-2 fuzzy neural network: Singular value decomposition approach," *Applied Soft Computing*, vol. 105, 107244, 2021.

[32] D. You, X. Gao, and S. Katayama, "WPD-PCA-Based Laser Welding Process Monitoring and Defects Diagnosis by Using FNN and SVM," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 628-636, 2015.

[33] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24-49, 2021.

[34] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, "CNN-Based Projected Gradient Descent for Consistent CT Image Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1440-1453, 2018.

[35] T. Zhang, D.-G. Zhang, H.-R. Yan, J.-N. Qiu, and J.-X. Gao, "A new method of data missing estimation with FNN-based tensor heterogeneous ensemble learning for internet of vehicle," *Neurocomputing*, vol. 420, pp. 98-110, 2021.