www.cad-journal.net
Computer-Aided Design AND APPLICATIONS™

Music Classification and Identification Based on Convolutional Neural Network

Yang Yuan[1] and Jiaqi Liu[2]

[1]School of Arts and Physical Education, Zhengzhou Vocational University of Information Technology, Zhengzhou 450008, China, ysys2018yuanyang@126.com
[2]Zhengzhou Normal University, Zhengzhou 450044, China, liujiaqi@zznu.edu.cn

Corresponding author: Jiaqi Liu, liujiaqi@zznu.edu.cn

Abstract. In this article, theoretical analysis and empirical research are combined. At first, the basic principles and applications of CAD and CNN (Convolutional Neural Network) are introduced. Then, how to apply these two technologies to music classification and identification is elaborated in detail, and a modelling and fusion method of music classification and identification is designed. Experiments show that the proposed innovative method is effective in music classification and identification. The performance of the method is evaluated by simulation experiments on different types and styles of music data sets and compared and analyzed with traditional music classification methods. The results show that, compared with the traditional music classification methods, the method proposed in this article can significantly improve the accuracy of music classification under the condition of limited labeled data, And its response speed is fast. The results fully prove the superiority of this method in classification accuracy and provide a new solution for MIR (Music Information Retrieval), recommendation, and other applications.

Keywords: Computer-Aided Design; Convolutional Neural Network; Music Classification; Music Identification
DOI: https://doi.org/10.14733/cadaps.2024.S18.205-221

1 INTRODUCTION

In the current digital age, music has become a part of people's daily lives, and the emergence of a large quantity of music works makes music classification and identification an important research field. The application of deep learning technology in music recognition and speech processing has made significant progress. For music recognition, deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) can effectively extract features from audio signals, thereby achieving tasks such as music classification, retrieval, and annotation. In terms of speech processing, deep learning technology can also be used in fields such as speech recognition, speech synthesis, and speech sentiment analysis. Bogach et al. [1] preprocessed the collected audio

using deep learning models. This includes steps such as converting audio into waveform maps and extracting audio features. Next, we will use another deep learning model to evaluate the learner's pronunciation. This model can be a convolutional neural network or a recurrent neural network, which can automatically identify errors and inaccuracies in learners' pronunciation. Finally, we provide feedback and suggestions based on the evaluation results. Feedback can include aspects such as accuracy of pronunciation, intonation, and speaking speed, while suggestions can include suggestions on how to improve pronunciation and practice methods. Due to the rapid growth of computer technology, artificial intelligence and signal processing technology, music classification and identification methods based on machine learning (ML) have been widely discussed. Music is a universal form of art that can inspire people's emotions and provide a joyful experience. However, for visually impaired individuals, they may not be able to appreciate music in traditional ways. To solve this problem, we can use CAD technology and Convolutional Neural Networks (CNN) to provide a new way of music appreciation, conveying music information through visual stimuli. Cho [2] explored how to implement this process and studied its impact on the multisensory experience and music recognition of visually impaired individuals. CAD technology can be used to extract structural information of music, such as notes, rhythm, and chords, and convert this information into visual expression. For example, notes and rhythms can be transformed into different colours and shapes, while chords can be transformed into different textures and images. These visual expressions can be conveyed to visually impaired individuals through special visual devices such as closed-circuit television or virtual reality glasses. Convolutional neural networks can be used to process these visual expressions and recognize information such as music types, styles, and emotions. By training neural networks to learn and classify different types, styles, and emotions of music, we can help visually impaired individuals better understand and appreciate music. In particular, as a branch of deep learning (DL), CNN has shown great potential in music classification and identification because of its excellent feature learning and classification capabilities. With the rapid development of information technology, music education is undergoing a revolution. Music education is no longer limited to traditional classroom teaching but expands to a wider range through digital and networked methods. Fu et al. [3] explored how to use the Analytic Hierarchy Process (AHP) and TOPSIS (Ideal Solution) to conduct decision analysis on music education based on information technology. In music education, we can use AHP to evaluate various information technologies, such as artificial intelligence, virtual reality, big data, etc., to determine their advantages, disadvantages, and applicability in music education. In music education, we can use TOPSIS to evaluate the performance of various information technologies, such as teaching effectiveness, learning efficiency, user satisfaction, etc., to determine their advantages, disadvantages, and applicability in music education. The music education decision-making model based on AHP and TOPSIS can be achieved through the following steps: first, using AHP to evaluate various information technologies and obtain the weights of each technology; Secondly, use TOPSIS to evaluate the performance of each technology and determine the relative advantages and disadvantages of each technology; Finally, combining the results of AHP and TOPSIS, the final decision result is obtained.

As an important branch of computer science, CAD has been widely used in many fields, such as architecture, machinery, and electronics. Music information visualization is a technique that converts music data into visual expression, which can help us better understand and analyze the structure and features of music. Classical composition is a traditional way of music creation that emphasizes the harmony and beauty of music. Georges and Seckin [4] explored how CAD technology and Convolutional Neural Networks (CNN) can be applied to music information visualization and classical composition to achieve music classification and recognition. Music information visualization can convert music data into graphics or images through CAD technology. For example, notes can be converted into different colours or shapes, while chords can be converted into different textures or images. These graphics or images can be transmitted to the audience through visual devices such as closed-circuit television or virtual reality glasses. Through this approach, visually impaired individuals can appreciate and understand music, while ordinary people can enhance their music experience through visual effects. Convolutional neural networks can be used to process and analyze classical music works and identify information such as music types, styles, and emotions. By training neural

networks to learn and classify different types, styles, and emotions of music, we can help people better understand and appreciate classical music. In recent years, with the continuous growth of computer technology, CAD has made remarkable progress in data processing, visualization and simulation. The traditional teaching method often focuses on imparting theoretical knowledge and skills while neglecting the cultivation of music perception ability. With the advancement of technology, the development of computer-aided teaching (CAD) and artificial intelligence has provided us with new solutions. The combination of CAD technology and Convolutional Neural Networks (CNN) especially provides new possibilities for training and improving music perception abilities. Gorbunova and Plotnikov [5] explored how to use CAD technology and CNN to achieve a multimodal teaching approach for music perception. By using CAD technology, teaching content such as music theory knowledge, skills, and performances can be digitized, displayed, and taught through multimedia devices. This approach can enhance students' understanding and mastery of music knowledge while also enhancing their interest and motivation in learning. By using CAD technology, teaching content such as music theory knowledge and skills can be transformed into digital audio files, allowing students to understand and master music knowledge while listening. Meanwhile, utilizing CNN for in-depth analysis of music audio data helps students perceive the features and patterns of music.

However, its application in music classification and identification is relatively rare. Music note recognition is the foundation of tasks such as music information retrieval, music classification, and music synthesis. Traditional music note recognition methods are usually based on rules or statistical models, which require a large number of manually designed features and have poor performance in recognizing music notes in complex and noisy situations. With the development of deep learning technology, the successful application of Convolutional Neural Networks (CNN) in fields such as image and speech has also promoted its application in music note recognition. He and Ferguson [6] proposed a music note recognition method based on CAD technology and convolutional neural network segmented two-stage learning. CAD technology can extract structural information of music, such as notes, rhythm, etc., which can be used to construct feature representations of music. In music note recognition, CAD technology can be used to convert music signals into a form suitable for convolutional neural network processing. For example, note and rhythm information can be transformed into images or feature vectors and then input into convolutional neural networks for processing. In music note recognition, segmented two-stage learning can be used to segment music signals, extract feature representations of each paragraph, and then use these feature representations for note classification or recognition. CAD has powerful data processing and visualization capabilities, which can provide new ideas and methods for music classification and identification. Music hearing loss is a common sensory defect that affects people's quality of life. Traditional music listening testing methods often rely on manual operation, which has subjectivity and significant errors. With the development of computer technology and artificial intelligence, auditory-based music hearing loss prediction has gradually become a research hotspot. Ilyas et al. [7] explored how to use computer-aided design and neural network tools to achieve automation and intelligence in music hearing loss prediction based on auditory perception. In the prediction of music hearing loss, CAD technology can be used to construct mathematical models of the ear canal, simulate the propagation process of sound waves of different frequencies in the ear canal, and thus obtain parameters that reflect hearing loss. Specifically, it is necessary to first obtain the structural data of the ear canal through medical imaging technology and then use CAD technology to convert this data into a three-dimensional model. Next, dynamic analysis of the model was conducted using methods such as finite element analysis (FEA) to obtain parameters that reflect hearing loss. Through this approach, quantitative assessment and prediction of music hearing loss can be achieved. Therefore, this article aims to explore the combination and application of CAD and CNN in music classification and identification.

The purpose of this article is to use CAD and CNN to realize music classification and identification, solve the problems existing in traditional music classification and identification methods, and improve the accuracy of music classification and identification. This article introduces a series of innovations in the field of music classification and identification, including multi-scale feature extraction, transfer

learning application, fusion model design, experimental verification and performance evaluation. These innovations not only improve the accuracy of music classification and identification but also provide new ideas and methods for research and application in related fields. The details are as follows:

⊖ Traditional music classification methods usually use a single-scale feature extraction method, but this article proposes a music classification method based on MSCNN (Multi-scale Convective Neural Networks). By constructing neural network models with convolution kernels of different scales, the multi-scale features of music data can be extracted so as to capture the internal structure and information of music more comprehensively. This method can improve the accuracy and robustness of classification and has good adaptability to different types and styles of music data.

⊖ Aiming at the problem of music classification with limited labeled data, this article introduces transfer learning technology and proposes a music classification method based on CNN and transfer learning. By using the CNN model pre-trained on large-scale image data sets and combining it with transfer learning technology, the pre-trained model is applied to music classification tasks. This method can solve the problem of limited annotation data and improve the accuracy of music classification.

⊛ This article proposes a music emotion identification method combining CNN and recurrent neural networks. By combining audio signal processing and DL technology, a fusion model is constructed, and the emotion identification and classification of music works are realized. This method combines the advantages of CNN in feature extraction and circular neural networks in sequence modelling, improves the accuracy of music emotion identification, and has a specific generalization ability for different types and styles of music data.

This article first introduces the basic principles and applications of CAD and CNN. Then, it expounds on how to apply these two technologies to music classification and identification. Then, a modelling and fusion method for the music classification identification problem is designed. Finally, the effectiveness of the proposed innovative method in music classification and identification is verified by experiments.

2 RELATED WORK

The digital music market is rapidly developing. Consumers can access music content anytime, anywhere, through various devices. Meanwhile, the application of CAD technology and Convolutional Neural Networks (CNN) has also played an important role in the field of digital music. Im et al. [8] explored the impact of CAD technology and CNN media services on the concentration of digital music consumption. In the field of digital music, CAD technology is widely used in music production, editing, and mixing. It can digitize the traditional music production process and improve production efficiency and quality. User preferences are one of the important factors affecting the concentration of digital music consumption. Users can choose their favourite music genres and singers according to their preferences, forming different consumer groups. The strategy of digital music platforms will also affect consumer concentration. The platform can guide users to consume specific music content through recommendation algorithms, personalized recommendations, and other methods, thereby affecting consumption concentration. CAD technology and CNN media services can improve the user experience of digital music platforms, thereby increasing user stickiness and consumption concentration. For example, through CNN's recommendation algorithm, users can be recommended their favourite music content, improving user experience and satisfaction. In digital media processing, audio compression technology is an important technique that can effectively reduce the size of audio files for storage and transmission. However, audio compression often brings certain sound quality losses. To address this issue, Lattner and Nistal [9] consider using Generative Adversarial Networks (GANs) to recover recompressed music audio. Generative Adversarial Network is a deep learning model consisting of two parts: a generator and a discriminator. The task of the generator is to generate new data, while the task of the discriminator is to determine whether these new data are true. During the training process, the generator attempts to deceive the discriminator

while the discriminator attempts to maintain vigilance against the generator. This process of mutual confrontation continuously enhances the generator's generation ability. For the recovery of recompressed music audio, we can design a GAN model, where the generator is responsible for learning and generating uncompressed audio from the original audio, and the discriminator is responsible for determining whether the generated audio is authentic. In this way, we can utilize the generative power of GAN to recover audio that has been recompressed. With the popularization of digital music and the rapid growth of online music, music information retrieval and fuzzy search have become increasingly important. Users often hope to find music that suits their preferences through fuzzy search or to understand the characteristics of specific music through information retrieval. The combination of CAD technology and Convolutional Neural Networks (CNN) provides a new solution for music information retrieval and fuzzy search. Lee and Hu [10] discussed the application and development of CAD technology and convolutional neural networks in music information retrieval and fuzzy search. CAD technology can help us extract structural information of music, such as notes, rhythm, and chords, which can be used to construct feature representations of music. Convolutional neural networks can process these features and perform classification or similarity calculations on them. In music information retrieval, we can encode music features using convolutional neural networks and use these encodings to retrieve works similar to specific music. Fuzzy search is a way to search for relevant information based on partial or fuzzy information. In the field of music, users often can only provide partial or vague information, such as "light," "melancholic," "classical", etc. CAD technology and convolutional neural networks can help us transform this fuzzy information into specific music feature representations, thereby achieving a fuzzy search of music.

With the continuous development of digital music technology, the amount of music data is experiencing explosive growth. Effectively classifying and managing massive music data has become an important issue. The classification of music genres can not only help us better organize and search for music but also help us understand and appreciate the connotations of different music cultures. Liu [11] proposed an automatic classification method for various music genres based on a combination of CAD (computer-aided design) neural networks and intelligent algorithms. CAD neural network is a deep learning-based model that learns and recognizes complex patterns by simulating the workings of human brain neurons. In the problem of music genre classification, we can convert music signals into formats suitable for neural network processing and then use CAD neural networks for training and prediction. Through training, CAD neural networks can learn the features and patterns of different music genres, thereby achieving automatic classification of new music. With the development of digital media technology, the quantity and variety of music data are exploding, and how to effectively discover, understand, and recommend music has become an important issue. Traditional music recommendation systems mainly rely on the user's historical behaviour and music attributes, such as popularity and style, while ignoring the emotional aspect of music. Therefore, Melchiorre et al. [12] proposed an emotion-aware music tower block (EmoMTB) that combines computer-aided design (CAD) technology and convolutional neural networks (CNN) to achieve emotional analysis and recommendation of music. After extracting the emotional features of music, we use Convolutional Neural Networks (CNN) for deep learning of these features. CNN can effectively process multimedia data such as images and audio, extracting feature representations of music from them. This feature representation can capture the complex patterns and structures of music, providing strong support for music recommendation. The performance of EmoMTB was evaluated through experiments using publicly available music datasets. The experimental results show that EmoMTB can effectively recognize and extract emotional features of music and use these features for music recommendation. Compared with traditional recommendation systems, EmoMTB has significantly improved recommendation accuracy and user satisfaction.

With the rapid development of the intelligent Internet of Things, the sharing and transmission of music data in various devices and applications has become increasingly common. In order to better manage and recommend music content, music classification has become an important task. Traditional music classification methods are mainly based on manually extracted music features, such as spectral features, Mel frequency cepstral coefficients, etc. However, the extraction process of these features often requires a lot of manual intervention, and it is difficult to capture all the

information about music. In recent years, the rise of deep learning technology has provided new solutions for music classification. Convolutional neural networks (CNNs), especially due to their powerful feature extraction ability and generalization performance, have been widely used in fields such as image, speech, and natural language processing. Seo and Huh [13] discussed how to combine CAD technology with convolutional neural networks to achieve more accurate and efficient automatic music classification. To verify the effectiveness of the music automatic classification scheme based on CAD technology and convolutional neural networks, we conducted a series of experiments. In the experiment, publicly available music datasets were used for training and testing, including different categories of music styles, instruments, and emotions. By comparing different classification methods, we found that the scheme based on CAD technology and convolutional neural networks outperforms traditional music classification methods in terms of accuracy and robustness. Music classification is an important task in music information processing, which involves recognition of music style, judgment of emotions, classification of composition style, and so on. Traditional music classification methods are mainly based on manually extracted music features, such as spectral features, Mel frequency cepstral coefficients (MFCC), etc. However, the extraction process of these features often requires a lot of manual intervention, and it is difficult to capture all the information about music. With the development of deep learning technology, especially the widespread application of Convolutional Neural Networks (CNNs), we have the opportunity to construct a more efficient and accurate music classification model. Singh [14] introduced an efficient deep neural network model for music classification, which is mainly designed based on convolutional neural networks and combines specific needs and characteristics in the music field. Convert audio signals into data formats suitable for neural network processing. Then, these features will be processed through multiple convolutional layers, pooling layers, and fully connected layers. Convolutional layers can automatically extract local features of audio signals, while pooling layers can reduce the dimensionality of these features and reduce computational complexity. The final fully connected layer maps the extracted features to the classification space and outputs the category to which the music belongs.

Instrument recognition refers to the process of identifying instrument types by analyzing audio signals. This has broad application value in practical applications, such as music classification, music education, music production, etc. Traditional instrument recognition methods are usually based on manually extracted features, such as Fourier transform, Mel frequency cepstral coefficients, etc. However, these methods often require manual intervention and adjustment and perform poorly in the face of complex and ever-changing music scenes. In recent years, the rise of deep learning technology has provided new solutions for instrument recognition. Deep convolutional neural networks (CNNs), especially due to their powerful feature extraction ability and generalization performance, have been widely used in fields such as image, speech, and natural language processing. Solanki and Pandey [15] explored how to use deep convolutional neural networks for instrument recognition. In instrument recognition, deep convolutional neural networks can be used to analyze audio signals and automatically learn features that can distinguish different instruments. The experimental results show that deep convolutional neural networks have achieved good performance in instrument recognition tasks, outperforming traditional feature extraction methods. In the process of music creation and production, multi-note fusion is an important technique that can bring richer and more complex expressions to music. However, the fusion of multiple notes is not an easy task, as it requires an effective method to handle and coordinate the relationships between different notes. In recent years, artificial neural networks (ANNs) have been widely applied in the field of music processing, providing us with such a method. Tian [16] introduced an intelligent music multi-note fusion method based on artificial neural networks. By utilizing the preprocessed music multi-note dataset, we can train our neural network model. During the training process, we use optimization algorithms such as gradient descent to adjust the model parameters and minimize prediction errors. In addition, we can also use regularization techniques to prevent overfitting problems. After experimental verification, we found that the intelligent fusion method of music multiple notes based on artificial neural networks can effectively improve the efficiency and quality of music creation. Specifically, our method achieved high accuracy and excellent performance indicators in multi-note

fusion tasks. In addition, we also found that this method has good generalization ability and can adapt to different music types and styles.

In the field of music information processing, Convolutional Neural Networks (CNNs) have been widely used in various tasks, such as music classification, style transfer, and emotion recognition. In these applications, the selection of activation functions has a significant impact on the performance of the model. Wang et al. [17] investigated the impact of different activation functions on music recognition in a convolutional neural network model for singer expression recognition. In the task of singer expression recognition, we need to identify the emotional expressions of the singer during the singing process. This requires the model to be able to learn the emotional features of music and map these features to the emotional space through activation functions. Choosing the appropriate activation function can enhance the model's perception ability of music's emotional features and improve the classification accuracy of the model. For example, the ReLU activation function has good performance in handling music emotion classification tasks, as it can effectively suppress overfitting and improve the robustness of the model. This article conducted experiments on a publicly available dataset of singer-singing expressions and compared the classification accuracy, robustness, and computational complexity of various models. The experimental results show that the ReLU activation function performs well in music emotion classification tasks, with better classification accuracy and robustness than other activation functions. In the field of music information retrieval and music classification, note recognition is an important task. Accurately identifying notes in music can provide rich musical information and help deepen understanding of the structure and style of music. However, traditional note-recognition methods mainly rely on the processing and analysis of audio signals, often ignoring the contextual information of speech. Xu et al. [18] proposed a novel dual-mode note recognition algorithm that integrates mixed features of music signals and speech context, aiming to improve the accuracy and robustness of note recognition. In music, speech context information is equally important for note recognition. We use speech recognition technology to extract contextual features of speech. Specifically, we use pre-trained deep learning models to automatically annotate speech signals and then extract linguistic features, such as phonemes, syllables, words, etc., from the annotated results. These features can capture prosodic and semantic information in speech signals. Next, we will integrate music signal features and speech context features. We use an attention mechanism to achieve the fusion of two features. The attention mechanism can automatically learn the weight allocation between different features, thereby achieving collaborative processing of music signals and speech context.

Music is an important component of human culture, carrying rich emotions and meanings. In the process of music creation, the coordination of melodies is a key factor in determining the quality of music. However, the coordination of notes often depends on the music literacy and experience of the creators, so achieving automatic note coordination has become a challenging issue. Zeng and Lau [19] proposed a structured representation method for music melody sequences based on computer-aided design (CAD) and reinforcement learning, which can achieve automatic coordination of notes. The sequencer in music production software uses CAD technology to arrange notes. In addition, reinforcement learning has also been applied in the field of music generation, such as automatic composition and arrangement. However, how to combine CAD and reinforcement learning to achieve automatic coordination of notes is still an unresolved issue. Use CAD technology to represent music melody sequences structurally. Specifically, we first perform spectral analysis on music melodies to extract features such as pitch, intensity, and rhythm and then use these features to construct a high-dimensional space. In this space, the coordinates of each dimension represent the characteristics of a note, while the entire space represents the structure of the musical melody. Music note recognition is one of the important tasks in the fields of music information retrieval and automatic composition. Accurately identifying notes in music can provide rich musical information and help deepen understanding of the structure and style of music. However, due to the complexity and diversity of music, music note recognition is a challenging problem. Traditional music note recognition methods often rely on a single feature or model, making it difficult to achieve accurate and robust note recognition. Zhang [20] proposed a music note recognition method based on multi-feature fusion, which improves the accuracy and robustness of note recognition by fusing

multiple features. Multi-feature fusion is a widely used technology in the fields of machine learning and pattern recognition, which can combine the advantages of multiple features to improve classification or recognition performance. In the field of music note recognition, some research has explored the fusion of various features, such as audio features, spectrogram features, temporal features, etc. These features reflect the information of music signals in different aspects, such as pitch, intensity, rhythm, etc. However, how to effectively integrate these features to achieve more accurate note recognition is still a problem that needs to be solved.

3 CAD AND MUSIC CLASSIFICATION AND IDENTIFICATION

3.1 Modelling of Music Classification and Identification

Although some progress has been made in music classification and identification technology, there are still some problems and challenges that need to be further studied and solved. On the one hand, the existing music classification and identification methods still have some difficulties in dealing with complex music types and styles. How to improve the accuracy and robustness of the algorithm is an important research direction. On the other hand, combining advanced DL technology with traditional signal processing technology and ML algorithms to further improve the performance of music classification and identification is also a problem worth studying. In addition, with the continuous growth and complexity of music data, how to effectively process and analyze a large quantity of music data is also an important research trend. Using CAD to process and visualize music data can provide new ideas and methods to solve this problem. At the same time, combining CAD with DL technology to achieve more efficient and accurate music classification and identification is also a challenging research direction. This article will make an in-depth study on this. Music classification and identification is an important research direction in the field of MIR. Traditional music classification and identification methods are mainly based on signal processing technology and ML algorithms, such as methods based on audio feature extraction and classifiers. These methods have made some achievements in music classification and identification, but there are still some problems. For example, the identification effect of complex music types and styles is not good, and the robustness to noise and variation is insufficient. In recent years, the rapid growth of DL technology has provided new ideas and methods for music classification and identification. The music classification and identification method based on DL can use a large quantity of music data for training and learning, automatically extract the deep features of music, and achieve better classification and identification results. CAD is a computer-based computing, modelling and simulation technology which is widely used in many fields, such as architecture, machinery and electronics. Its basic principle is to design and model with computer software and then simulate and visualize with computer hardware. In the field of music, CAD is mainly applied to the identification and transformation of music notation, the editing and production of music works and so on. Specifically, CAD can convert music data into visual graphics or images so as to better understand and analyze the structure and characteristics of musical works.

The core of music classification and identification is to extract and classify music data effectively. In this process, it is necessary to convert music data into a format that can be processed by ML algorithms. Usually, this format can be an audio feature vector, spectrogram Mel spectrum, etc. In order to make better use of CAD to classify and identify music, it is necessary to preprocess music data and extract features. Pretreatment steps include audio file reading, format conversion and noise removal. In the feature extraction step, CAD can be used to visualize audio signals and extract feature vectors. Specifically, CAD can be used to convert audio signals into image forms such as spectrogram or Mel spectrum, and then image processing technology can be used to extract feature vectors.

3.2 Application Method of CAD in Music Classification and Identification

It is necessary to design an effective classification algorithm to apply CAD to music classification and identification. This article uses CNN for training and classification. Firstly, the music data, after preprocessing and feature extraction, are input into the CNN model for training and learning. Then, the trained CNN model is used to classify and identify the new music data. In this process, we need to pay attention to some problems. For example, how to choose the appropriate CNN model structure, how to set superparameters, how to optimize the model and so on. In addition, we need to consider how to combine CAD with the CNN model to achieve more efficient and accurate music classification and identification. The overall framework of music style identification constructed in this article is shown in Figure 1.

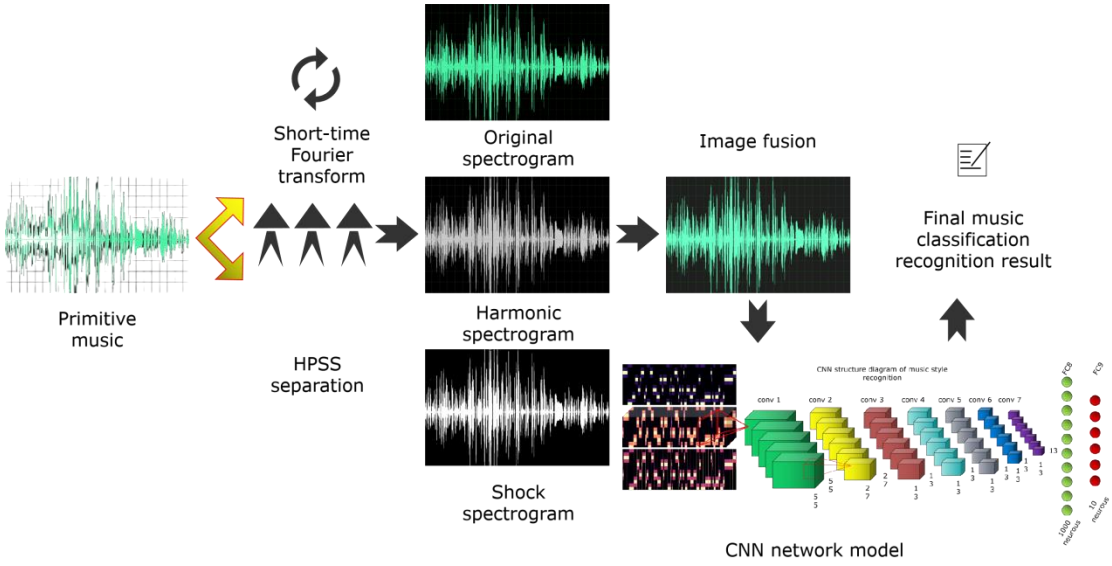

Figure 1: Overall frame diagram of music style identification.

4 CNN AND MUSIC CLASSIFICATION AND IDENTIFICATION

4.1 CNN Principle and Application

CNN is a DL model, which is mainly used to process image data. It has excellent performance in processing image data, so it is widely used in many fields. In the field of music, CNN has also been applied to music classification and identification. The basic structure of CNN includes a convolution layer, pooling layer, and fully connected layer. The function relation expression of the CNN model is:

$$y_i = sign(\sum_i w_{ji} x_i - \theta_i) \tag{1}$$

Among them, the input and output are both binary quantities W_{ji} with a fixed weight. After determining the network weights and parameters, the output of CNN can be expressed as:

$$y_i = \sum_{i=1}^{h} w_{ij} \exp(-\frac{\left\| x_p - c_i \right\|^2}{2\sigma^2}), j = 1,2,3,...,n \tag{2}$$

Where w_{ij} is the weight of the hidden layer and output.

In this article, CNN is used to learn and classify the characteristics of audio signals so as to realize the automatic classification and identification of different types and styles of music data.

4.2 Modelling the Problem of Music Classification and Identification

The modelling process of music classification identification mainly includes three steps: data preprocessing, feature extraction and classifier design. In the data preprocessing stage, audio files need format conversion, framing and normalization to facilitate subsequent feature extraction and classification. In the feature extraction stage, audio signal processing technology is used to extract the short-time energy, short-time zero-crossing rate, spectrum centroid and other features of audio so as to better describe the characteristics of music data. In the design stage of the classifier, this article chooses the appropriate ML algorithm to train and classify.

In order to make better use of CNN for music classification and identification, it is necessary to convert audio data into an image form that can be processed by CNN. In this article, the spectrogram or Mel spectrum of the audio signal is used to represent it. In this way, the problem of music classification and identification can be transformed into the problem of image classification for processing.

4.3 Application Method of CNN in Music Classification and Identification

To apply CNN to music classification and identification, it is necessary to design an effective CNN model structure and choose appropriate training and optimization methods. In the aspect of model design, this article chooses the classic AlexNet model structure to improve and optimize so as to better handle the time series information in music data. The CNN structure of this article is shown in Figure 2.

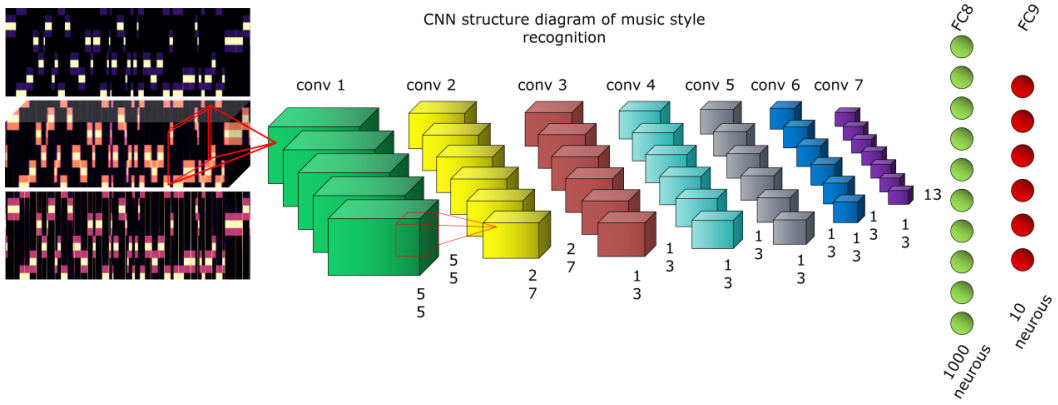

Figure 2: CNN structure.

The output of the last fully connected layer is the final identification result. This article uses Matlab to normalize the data. The formula is as follows:

$$f(x) = \frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}}(x - x_{\min}) + y_{\min} \tag{3}$$

Where $y_{\max}$ and $y_{\min}$ are the maximum and minimum values obtained by data transformation; $x_{\max}$ and $x_{\min}$ represent the maximum and minimum values of each sample.

Aiming at the problem of music classification with limited labelled data, this article proposes an innovative solution, that is, a music classification method based on CNN and transfer learning. The pre-training model is applied to the music classification task. In order to realize transfer learning, it is necessary to properly preprocess and extract features from music data. This can be achieved by converting the music data into a spectrogram, thereby converting the music data into a data format similar to an image. Input the preprocessed music data into the pre-trained CNN model for music

classification. In this process, this article uses fine-tuning technology to adapt to the task of music classification. Fine-tuning technology refers to training with limited labelled data on the basis of pre-training the model to adjust its parameters, thus improving the accuracy of music classification. Assume that the Fourier transform expression of the input sound signal is:

$$F(w) = S(w)H(w) \tag{4}$$

Where $S(w)$ is the Fourier transform of envelope information, and $H(w)$ is the Fourier transform of detail information? In the process of sound signal analysis and processing, we often pay more attention to the amplitude value, so it can also be expressed as:

$$\left|F(w)\right| = \left|S(w)\right|\left|H(w)\right| \tag{5}$$

Firstly, the logarithm operation can be performed on both sides $\left|F(w)\right|$ and then according to the property that the logarithm of the product of two numbers is equal to the logarithm sum of these two numbers, it can be transformed into a formula:

$$\log(\left|F(w)\right|) = \log(\left|S(w)\right|) + \log(\left|H(w)\right|) \tag{6}$$

The inverse Fourier transform of the above formula is shown in the following formula:

$$f(t) = s(t) + h(t) \tag{7}$$

In this way, the information on the envelope $s(t)$ and details $h(t)$ can be obtained.

Music data, especially audio data, contains information on various scales, such as rhythm, melody, harmony and so on. These different scales of information are very important for the classification of music. In order to extract this information more effectively, this article introduces a neural network model with convolution kernels of different scales. These convolution kernels with different sizes can capture the characteristics of music on different time and frequency scales so as to understand the internal structure of music more comprehensively. MSCNN consists of several parallel convolution paths, each of which has different convolution kernels. These paths can extract features at different scales and combine these multi-scale features through a fusion layer. This architecture allows the network to capture the music structure from local to global, which improves the network's ability to describe music. The model obtains the final identification result of the sample by averaging the identification results of each audio segment of the sample, and the calculation method is as follows:

$$Y_i = \begin{bmatrix} y_{i1}^{(1)} & y_{i1}^{(2)} & \cdots & y_{i1}^{(N)} \\ y_{i2}^{(1)} & y_{i2}^{(2)} & \cdots & y_{i2}^{(N)} \\ \cdots & \cdots & \cdots & \cdots \\ y_{iM}^{(1)} & y_{iM}^{(2)} & \cdots & y_{iM}^{(N)} \end{bmatrix} = \begin{Bmatrix} Y_{i1} \\ Y_{i2} \\ \cdots \\ Y_{iM} \end{Bmatrix} \tag{8}$$

Where $y_{ij}^{(K)}$ is the identification probability of the j audio segment of the i sample to the k class; Y_{ij} is the identification probability vector of all categories for the j audio segment of the i sample; Y_i is the identification probability matrix of all segments of the i sample for all categories, and H_i is the identification probability vector that the i sample belongs to all categories.

In order to train MSCNN, this article uses a large-scale music data set and adopts the standard DL optimization technology-back propagation. By optimizing the weights and parameters of the network, this article can make the network extract and classify the multi-scale features of music more effectively.

4.4 Experimental Design and Result Analysis

This section proposes a music classification method based on MSCNN and transfer learning. By using the CNN model pre-trained on large-scale image data sets and combining it with transfer learning

technology, the pre-trained model is applied to music classification tasks. This method can effectively solve the problem of limited annotation data and improve the accuracy of music classification. In order to verify its application effect in music classification and identification, experimental design and result analysis are needed.

In order to adjust the hyperparameters related to the model, this section randomly divides the data set into two subsets according to the ratio of 5:1. The results are shown in Figure 3.

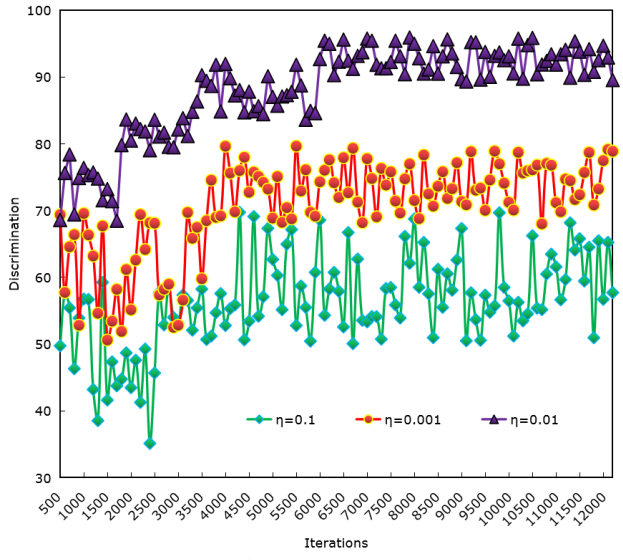

Figure 3: Learning rate η.

It can be seen that when the learning rate η is 0.001, the learning stage will be very slow, and the identification rate is not stable. If the learning rate η is 0.1, the learning stage will be unstable, and the classification performance will be reduced.

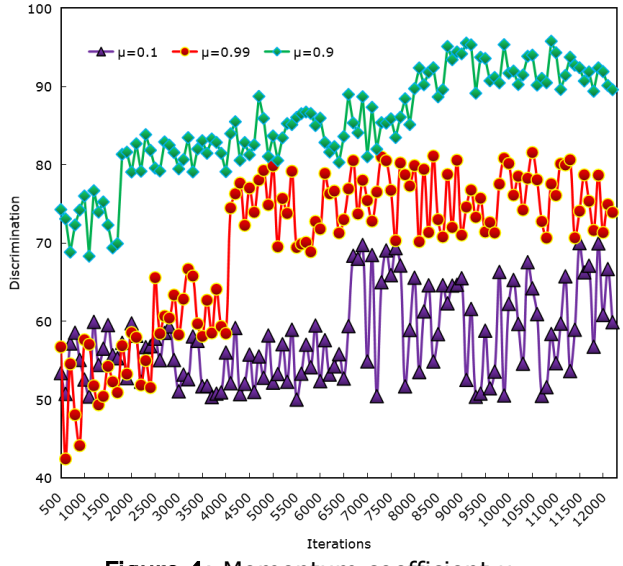

Figure 4: Momentum coefficient μ.

Figure 4 shows the momentum coefficient μ Experimental results. It can be seen that the use of momentum coefficient μ can speed up the learning stage well. In addition, it will reduce the classification performance in the later stage.

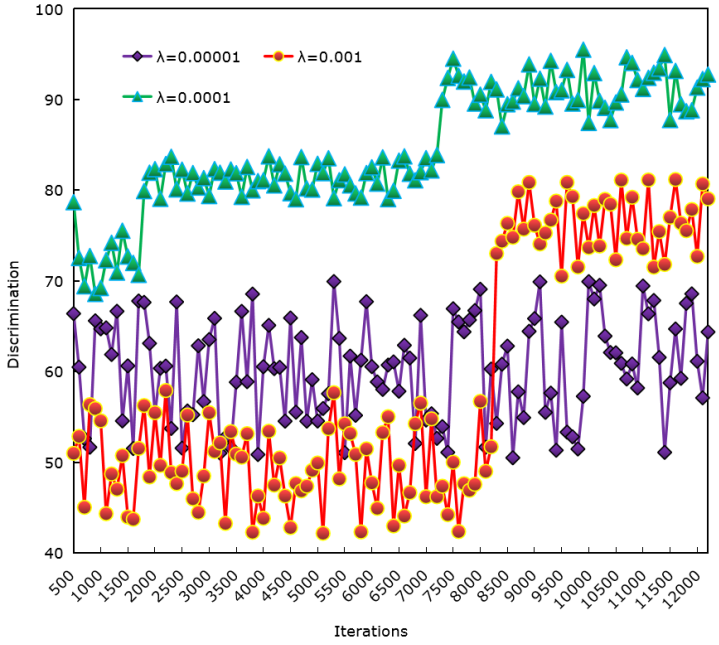

Figure 5: Weight attenuation coefficient λ.

Figure 5 shows the weight attenuation coefficient λ Experimental results. It can be seen that a smaller weight attenuation coefficient λ is a safer choice, while a larger weight attenuation coefficient λ will destroy the stability of the learning stage.

Through comprehensive analysis of the experimental results, this article sets the super-parameter as the best value in Table 1.

Hyperparameter	Set value
Learning rate η	0.01
Batch-size	15
Momentum coefficient μ	0.9
Weight attenuation coefficient λ	0.0001
Dropout coefficient	0.6

Table 1: Super-parameter setting value.

5 FUSION METHOD OF CAD AND CNN

5.1 Modelling and Fusion Method Design of Music Classification and Identification Problem

The fusion method of CAD and CNN is mainly based on their complementary advantages. CAD has powerful data processing and visualization capabilities and can convert music data into visual graphics or images so as to better understand and analyze the structure and characteristics of music

works. CNN is a DL model with excellent feature learning and classification ability, which can effectively extract the deep features of music data and classify and identify them.

The basic principle of integrating CAD and CNN is to preprocess and visualize music data through CAD and then input the processed data into the CNN model for feature learning and classification. In order to effectively integrate CAD and CNN into music classification and identification, it is necessary to design an appropriate fusion method. In this article, CAD is used to preprocess and visualize the audio data, and the audio signal is converted into the form of a Mel spectrum image. Mel frequency and Hertz frequency can be simply converted using the formula. A common conversion formula F_{mel} for converting f Hertz frequency into m Mel frequency is as follows:

$$F_{mel} = F_{mel}(F_{hz}) = 2595\log_{10}(1+\frac{F_{hz}}{700}) = 1127\log_{e}(1+\frac{F_{hz}}{700}) \tag{9}$$

Where F_{hz} is the frequency value of sound on the Hertz scale F_{mel}. It is the frequency value of sound under the Mel scale. Then, image processing technology can be used to extract and process the features of these images so as to better describe the characteristics of music data.

Next, we need to input the processed image data into the CNN model for feature learning and classification. In this process, it is necessary to select appropriate parameters such as CNN model structure, loss function, optimizer, and learning rate so as to train and optimize the model better. At the same time, we need to consider how to combine CAD with the CNN model to realize the effective integration of the two. In this article, transfer learning, data enhancement and other technologies are used to improve the performance and generalization ability of the model, and appropriate fusion strategies are selected to realize the organic combination of the two. In this way, we can make full use of the data processing and visualization capabilities of CAD and the feature learning and classification capabilities of CNN to achieve more efficient and accurate music classification and identification.

5.2 Experimental Design and Result Analysis

In order to verify the application effect of the fusion method of CAD and CNN in music classification and identification, experimental design and result analysis are needed. In this experiment, this method is compared and analyzed with other traditional music classification and identification methods.

This section selects different types and styles of music data sets for experimental verification and compares and analyzes them with other traditional music classification and identification methods. In the aspect of experimental design, this article determines the experimental purpose, data set selection, preprocessing and feature extraction methods, CNN model structure and super-parameter setting. In the process of experiment, we also need to pay attention to the confidentiality and repeatability of data. In the analysis of experimental results, this section uses visualization technology to show and analyze the experimental results. Figure 6 shows the experimental verification results of different types and styles of music data sets.

MSCNN can significantly improve the accuracy and robustness of classification compared with the traditional single-scale method. MSCNN has achieved better performance than the single-scale method on various music data sets, which proves the effectiveness of the multi-scale feature extraction method. The accuracy results of different music classification methods are shown in Figure 7.

Compared with traditional music classification methods, the method proposed in this article can significantly improve the accuracy of music classification with limited labelled data. The method in this article has achieved better performance than other methods in music genre and style classification tasks. The response speed index is used to evaluate and compare the classification performance, and the results are shown in Figure 8.

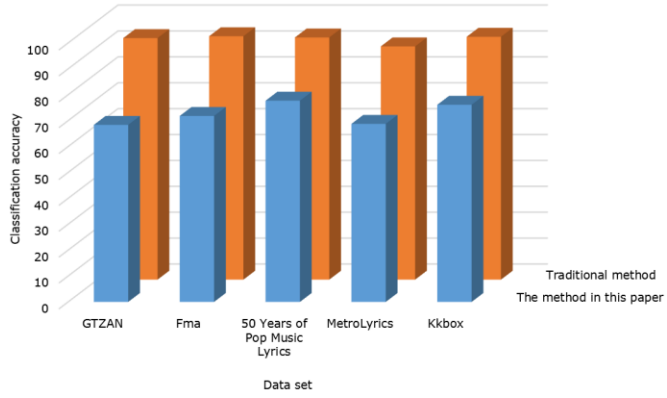

Figure 6: Experimental verification results.

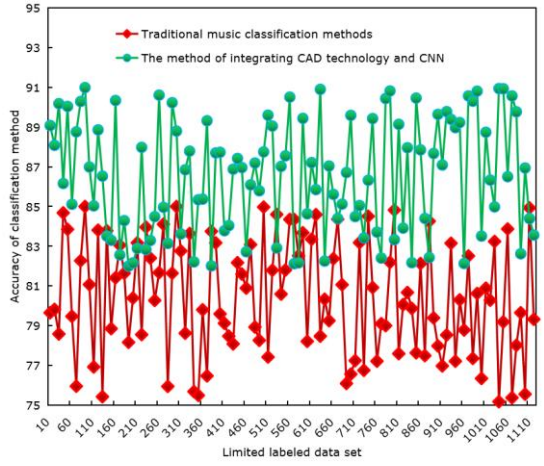

Figure 7: Accuracy of different music classification methods.

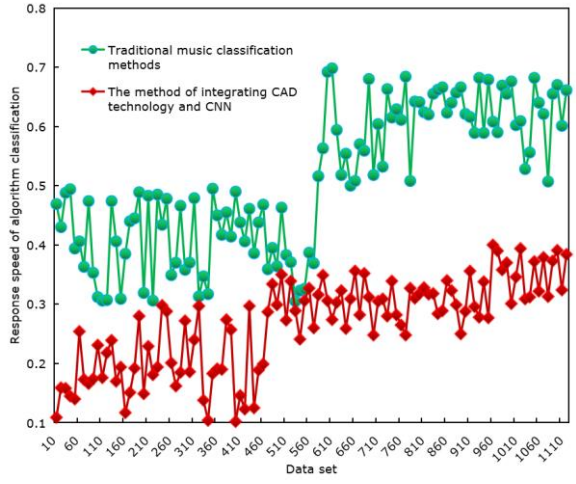

Figure 8: The response speed of algorithm classification.

After experimental verification and analysis, it is found that the method of combining CAD and CNN is excellent in music classification and identification. This method can not only effectively extract the deep features of music and classify and identify them but also show good adaptability and stability for various types and styles of music data. In addition, this method is expected to further improve the performance and efficiency of music classification and identification and provide better technical support and solutions for MIR, recommendation and other applications.

6 CONCLUSIONS

This article puts forward a music classification and identification method based on CAD and CNN, which realizes the automatic processing and classification of music data. Through experimental analysis, it is found that CNN can achieve better classification performance and effect. This method can effectively extract the deep features of music, classify and identify it, and has certain adaptability and robustness to different types and styles of music data. At the same time, this method can also be combined with other DL technologies to achieve more efficient and accurate music classification and identification. At the same time, we can explore more feature representation methods and model optimization strategies to improve the performance of music classification and identification further.

In this article, the feasibility and effectiveness of the proposed method are verified by simulation experiments, which provides new ideas and methods for future music classification and identification research. The research results can be applied not only to the music field, but also to data processing and classification in other fields, which has a wide application prospect. Future work can further explore how to optimize the pre-training model and transfer learning strategies to improve the performance of music classification. At the same time, we can also consider applying the method proposed in this article to other similar MIR and analysis tasks.

Yang Yuan, https://orcid.org/0009-0003-6402-6445
Jiaqi Liu, https://orcid.org/0009-0004-8509-3425

REFERENCES

[1] Bogach, N.; Boitsova, E.; Chernonog, S.; Lamtev, A.; Lesnichaya, M.; Lezhenin, I.; Blake, J.: Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching, Electronics, 10(3), 2021, 235. https://doi.org/10.3390/electronics10030235

[2] Cho, J.-D.: A study of multisensory experience and color recognition in visual arts appreciation of people with visual impairment, Electronics, 10(4), 2021, 470. https://doi.org/10.3390/electronics10040470

[3] Fu, Y.; Zhang, M.; Nawaz, M.; Ali, M.; Singh, A.: Information technology-based revolution in music education using AHP and TOPSIS, Soft Computing, 26(20), 2022, 10957-10970. https://doi.org/10.1007/s00500-022-07247-w

[4] Georges, P.; Seckin, A.: Music information visualization and classical composers discovery: an application of network graphs, multidimensional scaling, and support vector machines, Scientometrics, 127(5), 2022, 2277-2311. https://doi.org/10.1007/s11192-022-04331-8

[5] Gorbunova, I.-B.; Plotnikov, K.-Y.: Music computer technologies in education as a tool for implementing the polymodality of musical perception, Musical Art and Education, 8(1), 2020, 25-40. https://doi.org/10.31862/2309-1428-2020-8-1-25-40

[6] He, N.; Ferguson, S.: Music emotion recognition based on segment-level two-stage learning, International Journal of Multimedia Information Retrieval, 11(3), 2022, 383-394. https://doi.org/10.1007/s13735-022-00230-z

[7] Ilyas, M.; Othmani, A.; Naït, A.-A.: Computer-aided prediction of hearing loss based on auditory perception, Multimedia Tools and Applications, 79(21), 2020, 15765-15789. https://doi.org/10.1007/s11042-020-08910-w

[8] Im, H.; Song, H.; Jung, J.: The effect of streaming services on the concentration of digital music consumption, Information Technology & People, 33(1), 2020, 160-179. https://doi.org/10.1108/ITP-12-2017-0420
[9] Lattner, S.; Nistal, J.: Stochastic restoration of heavily compressed musical audio using generative adversarial networks, Electronics, 10(11), 2021, 1349. https://doi.org/10.3390/electronics10111349
[10] Lee, K.-Y.; Hu, C.-M.: Research on the development of music information retrieval and fuzzy search, Scientific and Social Research, 4(4), 2022, 1-10. https://doi.org/10.26689/ssr.v4i4.3771
[11] Liu, J.: An automatic classification method for multiple music genres by integrating emotions and intelligent algorithms, Applied Artificial Intelligence, 37(1), 2023, 2211458. https://doi.org/10.1080/08839514.2023.2211458
[12] Melchiorre, A.-B.; Penz, D.; Ganhör, C.; Lesota, O.; Fragoso, V.; Fritzl, F.; Schedl, M.: Emotion-aware music tower blocks (EmoMTB): an intelligent audiovisual interface for music discovery and recommendation, International Journal of Multimedia Information Retrieval, 12(1), 2023, 13. https://doi.org/10.1007/s13735-023-00275-8
[13] Seo, Y.-S.; Huh, J.-H.: Automatic emotion-based music classification for supporting intelligent IoT applications, Electronics, 8(2), 2019, 164. https://doi.org/10.3390/electronics8020164
[14] Singh, J.: An efficient deep neural network model for music classification, International Journal of Web Science, 3(3), 2022, 236-248. https://doi.org/10.1504/IJWS.2022.122991
[15] Solanki, A.; Pandey, S.: Music instrument recognition using deep convolutional neural networks, International Journal of Information Technology, 14(3), 2022, 1659-1668. https://doi.org/10.1007/s41870-019-00285-y
[16] Tian, Y.: Multi-note intelligent fusion method of music based on artificial neural network, International Journal of Arts and Technology, 13(1), 2021, 1-17. https://doi.org/10.1504/IJART.2021.115763
[17] Wang, Y.; Li, Y.; Song, Y.; Rong, X.: The influence of the activation function in a convolution neural network model of facial expression recognition, Applied Sciences, 10(5), 2020, 1897. https://doi.org/10.3390/app10051897
[18] Xu, Y.; Su, H.; Ma, G.; Liu, X.: A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context, Complex & Intelligent Systems, 9(1), 2023, 951-963. https://doi.org/10.1007/s40747-022-00841-3
[19] Zeng, T.; Lau, F.-C.: Automatic melody harmonization via reinforcement learning by exploring structured representations for melody sequences, Electronics, 10(20), 2021, 2469. https://doi.org/10.3390/electronics10202469
[20] Zhang, Y.: Music emotion recognition method based on multi feature fusion, International Journal of Arts and Technology, 14(1), 2022, 10-23. https://doi.org/10.1504/IJART.2022.122447