Research Article

Research on Music Content Identification and Recommendation Technology Based on Deep Learning

Ping Li

HeNan Open University, China
pingli1098@gmail.com

Received: 16th September 2024; Accepted: 25th December 2024; Published: 1st January 2025

Abstract: This research introduces a novel music recommendation system leveraging deep learning techniques to tackle significant challenges in traditional recommendation methods, such as the cold start problem, limited recommendation diversity, and difficulty in adapting to evolving user preferences. The proposed model employs Convolutional Neural Networks (CNNs) for genre recognition, coupled with Harmonic-Percussive Source Separation (HPSS) to extract rich audio features, capturing intricate musical distinctions across genres. These features, combined with user interaction data, enable the model to deliver highly personalized recommendations based on individual listening habits. Experimental results show that the system significantly outperforms conventional approaches, with a genre classification accuracy of 92%, offering greater recommendation accuracy and diversity. This marks a substantial improvement over traditional collaborative filtering and content-based methods, which struggle to deliver relevant suggestions in dynamic user environments. The findings highlight that deep learning, particularly CNNs, can effectively overcome data sparsity issues and provide more adaptive, user-centered recommendations. Moreover, the system's ability to integrate real-time user interaction data leads to enhanced user engagement, as the recommendations become more relevant and aligned with individual preferences. Future work will explore enhancing the dataset's diversity and optimizing computational efficiency to support scalability, ensuring the model can be applied across different cultures and regions. By improving the model's adaptability and efficiency, this research aims to create a more inclusive and scalable music recommendation system, capable of catering to global audiences with diverse musical tastes. Ultimately, the proposed system contributes to the development of more accurate, personalized, and engaging music recommendation frameworks, marking a significant advancement in the field of music information retrieval.

Keywords: ***Convolutional Neural Network; Deep Learning; Music Content Recognition; Recommendation Technology***

1. Introduction

As the Internet and technology continue to advance, digital music has become an integral part of everyday life. Thanks to the Internet's fast and efficient transmission capabilities, demand and supply for digital music have surged [1]. According to a 2016 report on the online music industry, the global digital music market reached $7.23 billion, marking a 7.6% increase from 2015. In China, since 2015, a series of standard digital music policies by the National Copyright Administration have helped the industry steadily develop. However, as digital music databases expand rapidly, traditional music retrieval methods are increasingly revealing their limitations [2-3].

On the one hand, the user retrieval needs are fuzzy, and the massive amount of information will produce the problem of insufficient vocabulary and information anxiety. The current retrieval method increases the difficulty of user retrieval [4].

In the face of the above problems, it is necessary to actively provide users with music services that align with their interests, and the music recommendation system came into being. Among the current major

Ping Li, "Research on Music content identification and recommendation technology Based on Deep Learning", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, Vol. 9, No. 1, pp. 58-72, 1st January 2025, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2025.01.004, Available: http://aetic.theiaer.org/archive/v9/v9n1/p4.html.

music websites. There are two recommendation systems: the list of recommendations, usic sites based on public listening habits, and Gathering some more popular tracks into music charts. On this basis, this can be recommended to users in a certain probability [5]. However, it cannot be personalized and recommended according to their needs. The other is the ability to customize different music for users; usually, based on information such as music content, user behavior, and user data, recommend users to music that may be of interest; although this method has some problems, such as the recommendation results. Still, it reflects the effect of personalized recommendation system recommendation systems. Current research in this direction is a hot topic. There is a great space for development [6]. In recent years, music recommendation systems became important tools in enhancing user experience on streaming platforms; however, their effectiveness strongly relies on both dataset diversity and computational scalability. Most of the current models, even those using advanced techniques such as Convolutional Neural Networks (CNNs) for deep learning, are trained predominantly on English-language datasets, which limits both their cultural relevance and their applicability across a global audience [7]. The different linguistic and cultural contexts in which people live influence music preferences to a great extent, and exploring datasets with genres and styles across regions would enrich the recommendation systems, capturing unique musical patterns and preference types. Additionally, the complexity in the computational requirements of deep learning models, especially for smaller platforms with fewer resources, might hinder wider implementation. These limitations should be addressed by investigating dataset diversity and model optimization techniques in developing a more universally applicable and accessible music recommendation system. This study recognizes the challenges of the present work and provides the ground for future work to build scalable, culturally inclusive recommendation frameworks that will leverage the strengths of deep learning while remaining practical in different resource environments.

Based on the identified limitations of the traditional music recommendation systems, this paper introduces a deep learning-based approach, using convolutional neural networks (CNNs) and deep confidence networks, to improve the recommendation precision and diversity. CNNs are used to recognize the style of the music in a spectrogram, allowing for the extraction of more complicated features concerning music in order to make better genre-specific recommendations. In addition, the multi-feature fusion method is used to integrate audio features, user behavior data, and historical preferences for a personalized recommendation system that caters to the taste of an individual.

The research process involves three main steps: first, analyzing existing recommendation methods to identify performance gaps; second, implementing the CNN-based model for effective music style classification and integrating it with a deep confidence network to combine relevant features; and third, evaluating the model's performance using key metrics such as accuracy, diversity, and adaptability. The proposed framework is built to deal with the cold start problem, enhance recommendation diversity, and provide a more solid, user-centered approach for modern music recommendation systems.

2. Review of Research at Home and Abroad

Researchers have made substantial advancements in algorithms for recommendation systems, leading to their widespread adoption across various fields. One of the most influential breakthroughs in this area was the collaborative filtering algorithm, which laid the foundation for modern recommendation systems [8]. This pivotal development paved the way for recommendation applications, including music. In 1995, the music recommendation field gained traction with the MIT laboratory's release of the Ringo system, designed to suggest engaging music content and predict user song ratings [9]. Ringo's rapid success set a high standard that other music recommendation systems struggled to surpass for years, as most relied on traditional music information, resulting in less innovative recommendations.

The development of music recommendation systems has continued to evolve, with major music platforms globally enhancing their recommendation technologies. International platforms like Last.fm and Pandora have pioneered distinct approaches. Last.fm leverages users' listening histories and preferences to connect individuals with similar tastes, thus curating music that aligns with shared interests. In contrast, Pandora relies on analyzing core musical characteristics to assess song similarities. According to recent study [10], an effective recommendation strategy should offer a wider selection by compiling multiple sets of highly recommended music into playlists, increasing the chance of user satisfaction. Additionally, other

researchers have explored audio modeling techniques to visually represent user preferences [11] and applied weighting to various musical elements based on user tastes [12].

In China, although research on music recommendation systems started relatively late, platforms like Douban, Xiami, and Netease Cloud Music have rapidly grown, integrating unique, socially-driven features. These platforms consider user behaviors—such as listening habits, social interactions, and music sharing—enabling more relevant music recommendations. Chinese researchers have also made notable strides in recommendation algorithms. For instance, an effective method of grouping music and utilizing user ratings through probability modeling was proposed [13], enhancing recommendation accuracy. Another study by Zhang Yan and colleagues employed fractal theory to reduce the complexity of music libraries, helping to alleviate storage strain on databases by reducing data dimensions.

A major challenge in music recommendation systems is data sparsity, as users generally engage with only a small portion of the extensive music database. This results in a sparse user-score matrix, reducing the accuracy of recommendation algorithms. To address this, researchers have developed a collaborative filtering recommendation algorithm based on music score prediction [14], which calculates score correlations to improve the user-score matrix. Additional methods have been proposed to mitigate sparsity, such as dividing users into groups for predictive scoring [15] and using BP neural networks [16] to estimate scores, partially resolving the sparse matrix problem. Other solutions include the K-means clustering collaborative filtering algorithm with SVD matrix filling, leveraging hidden user-item relationships to address sparsity [17]. Researchers have also explored various matrix-filling techniques, including mean filling, linear regression, and Bayesian classification, to compare and enhance the accuracy of recommendation predictions [18].

The cold start problem in recommendation systems occurs when there is no historical interaction data to make accurate recommendations for new users or new items. If a new song has just been added to a digital music platform and there is no prior user engagement with the song in the form of listens, ratings, or reviews, then there is not enough information to determine which users will likely be interested in it. Equally, the system struggles to recommend a new user personalized items with no history of interaction with any items. This makes it difficult for the traditional algorithms of recommendation to come up with relevant content and, as such, delivers the ability to deliver timely and personalized experiences. To address this issue, Liu [19] suggested a recommendation method that combines user preferences and music features. This approach helped to mitigate the impact of the cold start problem by reducing the reliance on user ratings for new music recommendations. Another proposed solution, presented in Kim *et al.*'s study [20], was a cold start recommendation method based on grain association rules. This method utilized grains to describe users and products, and by meeting the criteria of grain association rules, it uncovered association rules between users and products. These rules were then used to generate appropriate recommendations. Additionally, Zhang *et al.* [21] tackled the cold start problem by leveraging the similarity of users' music evaluations and the correlation of music. By considering users' preferences, this method offered a recommendation approach that partially alleviates the cold start problem in the recommendation algorithm.

Hou's study [22] proposed a CNN-based music recommendation system that outperforms traditional methods by achieving 95% accuracy. This AI-driven system enhanced personalization, addressing challenges in large music libraries by analyzing content for tailored suggestions. The approach demonstrated superior efficiency and precision over other algorithms, including deep neural network (DNN), Recurrent Neural Network (RNN) , LSTM, and traditional models like SVM and KNN. Wen [23] developed an intelligent background music system using deep learning and IoT. Utilizing a novel middle-level feature extraction approach, the system achieved an 87.6% accuracy rate in recognizing indoor scenes, outperforming traditional methods, especially in varied lighting. Implemented in an Intelligent Home, the system proved stable and effective, laying groundwork for future smart music applications. Zhang's study [24] addressed the challenge of selecting preferred music from vast databases by developing a CNN-based recommendation system focused on digital piano music. It extracted spectrum and note features, refined classification results, and used user behavior to improve model accuracy. Two methods—single-category and multicategory recommendation—were tested, with multicategory features showing higher accuracy. The single-category method achieved 50.35% accuracy, while multicategory features improved recommendation precision.

Despite the development of knowledge graph-based music recommendation, the current model still fails to capture dynamic changes of users' preferences and diversity in music genres. Traditional approaches of collaborative filtering and content-based filtering almost always suffer from the sparsity of users' data and the 'cold start' problem of new users and items. Although deep learning techniques have shown to enhance the accuracy of recommendation, there is still room for improvement by more robust frameworks to deeply integrate diverse sets of audio features and users' interaction data so that the accuracy and personalization of recommendation are lifted. This paper narrows the above gap by proposing a deep learning-based recommendation model, taking the advantages of CNNs for music style recognition and deep user interaction data aiming at creating a system capable of managing dynamic and individualized preference.

A novel music recommendation approach based on deep learning techniques is proposed in this paper by utilizing a CNN-based model and several new preprocessing methods, including HPSS (Harmonic-Percussive Source Separation) and multi-feature fusion. This model will improve both accuracy and diversity in recommendation compared to traditional recommendation systems that cannot cope with data sparsity and the 'cold start' problem by focusing on complex audio feature extraction and the user's personalized interaction data. Existing methods, such as collaborative filtering and basic content-based filtering, are usually inadequate to model nuanced user preferences—especially in dynamic environments where user tastes change frequently. Proposed approach fills these gaps by providing a much stronger solution for making diverse and precise recommendations—something very important in today's enlarged music libraries. By those state-of-the-art techniques, this study overcomes not only the limitations of the previous models but also sets a new standard for adaptive, user-centered music recommendation systems, showing clear improvement in the relevance of recommendations and in user satisfaction.

The academic research on music recognition, recommendation algorithms, and deep learning is very broad. Still, deep learning is applied to identify music types, and recommendation technology is used to generate recommendation models based on music recognition, so this paper has certain experimental and theoretical significance.

3. Proposed Model for Music Recommendation

3.1. Deep Neural Network Theory

Although the traditional recommendation methods, such as collaborative and content-based filtering, work fairly well, they are usually unable to adapt to complicated user preferences and the immense variety of music nowadays. Deep learning appears to be a great potential solution for these problems since it can be used in extracting complex patterns within huge datasets, such as those found in music libraries. It can improve the understanding of subtle audio features and user behaviors in recommendation systems through deep neural networks, hence leading to more accurate and personalized music suggestions. This approach positions deep learning as a key advancement in addressing the evolving demands of music recommendation.

3.1.1. Cyclic Neural Network

RNNs are a type of neural network designed to handle sequential data where order matters—think text, audio, or time series data. The main difference between RNNs and feedforward neural networks is that RNNs have feed-back connections, allowing the retaining and passing of information from one step in the sequence to the next. The feedback mechanism allows the RNN to 'remember' past inputs, hence carrying out tasks involving context and temporal dependencies more effectively—for example, music recommendation, where past listening history might influence future recommendations.

This work employs a CNN architecture for capturing detailed audio features of music style recognition based on the Mayer spectrum maps as input to utilize the rich frequency and temporal information. Although effective in complex pattern recognition, CNNs are computationally costly and may not be readily available on smaller platforms due to the lack of resources. To address this, we look into optimization techniques that can curtail computational overhead without affecting model performance. Optimization techniques, from model pruning and quantization to lightweight architectures such as MobileNet and EfficientNet, provide promising paths to make the model more resource-efficient. These optimizations will

help ensure that the model can be run in an effective manner on platforms with reduced computational power, such as mobile devices or low-power servers, so as to extend applicability. Future work will be devoted to experimenting with these strategies to keep the model both robust and adaptable to different deployment environments, which will make the model even more scalable and practical.

At each time step, the input at that time is taken in by the RNN, and it combines that input with information it's seen in past steps, stored in its hidden state. It is then able to pick up patterns over time, making it very fitting in the recognition of song sequences or user behavior trends. However, in the standard RNNs, information over long sequences might get lost because of problems such as the vanishing gradient problem. More complex RNN architectures, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), introduce gating mechanisms through which the network can selectively retain or forget information, enabling the capture of longer-term dependencies.

The prototype of a RNN is first proposed in 1982 by American physicist John Hopfield [25]. The basic feature includes at least one feedback connection in the network, allowing the activation function to be repeated in the loop, thus allowing the network to realize the processing and learning ability of time series. Different from the standard Feed Forward Neural Network (FFNN), at the current moment, the RNN retains the hidden layer state of the previous moment, that is, remembering certain historical information, so when the input sequence is encoded into a vector, the structural features of the sequence can be retained. Fig. 1 shows the RNN and the structure expanded in a time step, representing the input, hidden, and output layers.

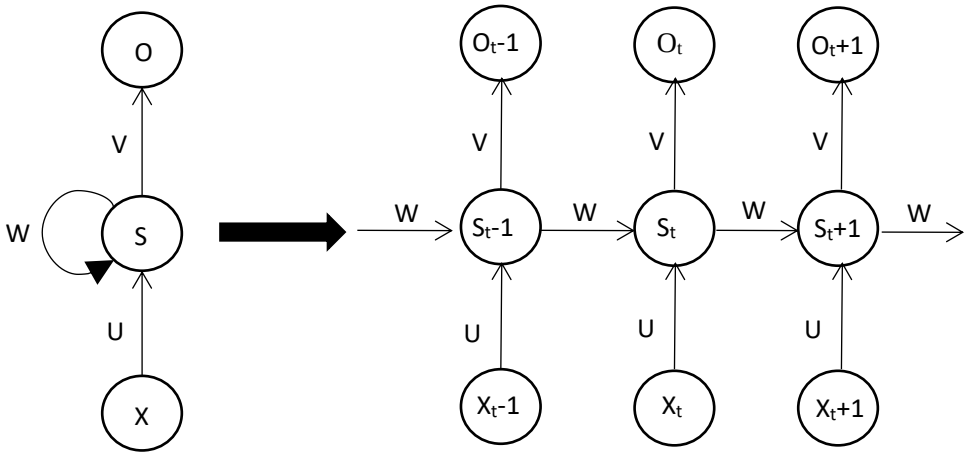

Figure 1. Structural diagram of RNN expanded by time steps

Assuming the input sequence X={x_1,x_2,.x_3 ... x_n} of the model, x_t represents the information input to the network at time t, x_t represents the hidden layer state of the network at time t, o_t represents the output of the network at time t. At each time step, s_t, the hidden state, is calculated as a function of the current input x_t and the previous hidden state s_{t-1} using Eq. (1). The output layer is predicted with $\hat{y}$, s_t and $\hat{y}$ are calculated as shown in Eqs. (1) and (2):

$$s_t = f(Ux_t + Ws_{t-1})_{tanh} \tag{1}$$

$$\hat{y} = g(Vs_t)_{softmax} \tag{2}$$

In which, U and W are weight matrices that help transform inputs and hidden states to calculate the new hidden state at each time step. The RNN model shows good results in the research fields such as text generation and machine translation. However, the model still has defects, such as gradient vanishing, gradient explosion, and long-distance dependence [26]. RNN usually uses the time-based backpropagation algorithm in training, and the chain conduction method is adopted in the guidance process. The final loss value is the sum of loss at all previous moments, as shown in Eq. (3):

$$\frac{\partial E}{\partial W} = \sum_t \frac{\partial E_t}{\partial W} \tag{3}$$

The unique feature of BPTT is that when W is guided, the chain guidance method is applied, and time t=3 is taken as an example; it is easy to get Eq. (4) and Eq. (5):

$$\frac{\partial E_3}{\partial W} = \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial s_3} \frac{\partial s_3}{\partial W} \tag{4}$$

$$s_3 = tanh(Ux_3 + Ws_2) \tag{5}$$

It can be seen that s_3 is dependent on s_2 and W, and continues to derive the gradient at t=3, as shown in Eq. (6):

$$\frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial s_3} \frac{\partial s_3}{\partial s_k} \frac{\partial s_k}{\partial W} \tag{6}$$

Due to the propagation of a gradient value throughout the network, the calculation process can be significantly impacted. When the gradient value becomes excessively large or small, it can lead to the problems of gradient disappearance or gradient explosion. Additionally, as the sentence length increases, the drawback of retaining long-distance information becomes more pronounced, significantly affecting the RNN's effectiveness. Researchers have developed two classical models to address this issue: the Long-Term and Short-Term Memory Network Unit and the Gating Cycle Unit. Both models incorporate gate mechanisms to selectively retain valuable historical information from previous moments, enabling the RNN to capture information over longer distances effectively.

While RNNs work well with sequential data and, by extension, temporal dependencies, they are less effective in extracting spatial patterns in data, such as those found in audio spectrograms or visual representations of sound. For music recommendation, features extracted from the audio in both the frequency and time-domain are used for genre and style classification. CNNs are good at capturing spatial hierarchies in such data, which makes them very suitable to analyze the complex patterns in music spectrograms. Introducing CNNs alongside RNNs allows the recommendation system to consider both temporal and spatial insights for a better comprehension of music content.

3.1.2. Convolutional Neural Network

In recent years, the widespread attention garnered by deep learning and the advancements in numerical computing equipment have greatly enhanced the representation learning capability of CNNs [27]. Consequently, one of the primary areas of research has been focused on CNNs, which are composed of multiple hidden layers, an input layer, and an output layer. Each layer comprises numerous neurons, as illustrated in Fig. 2.

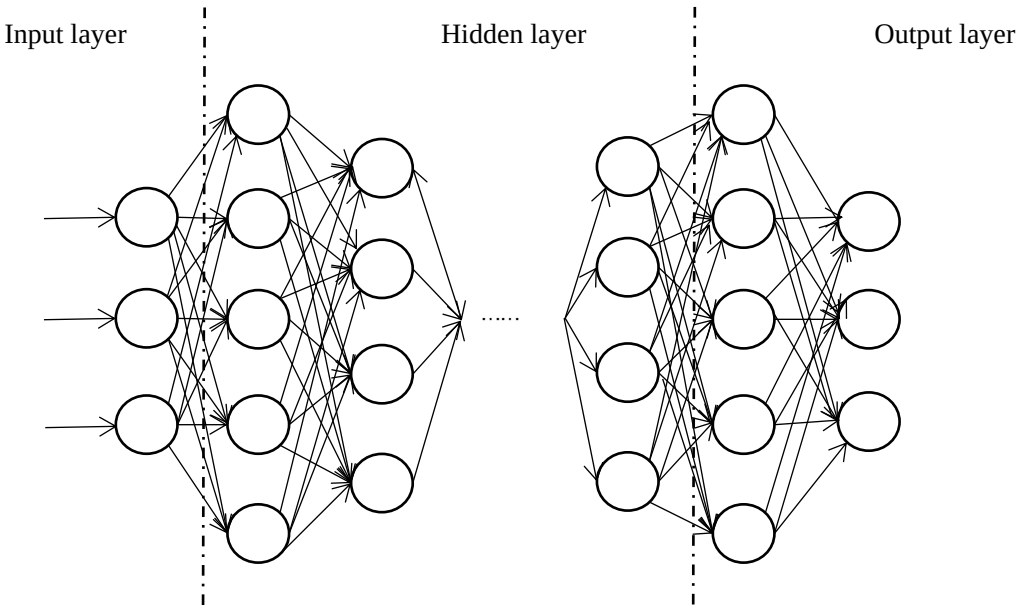

Figure 2. Schematic representation of the neural network

According to Fig. 2, the lower and upper layer neurons can form connections in the traditional fully connected deep neural network (DNN). This comprehensive series with multiple levels promptly highlights the issue of increasing the parameters, which can lead to overfitting and getting trapped in a local optimum.

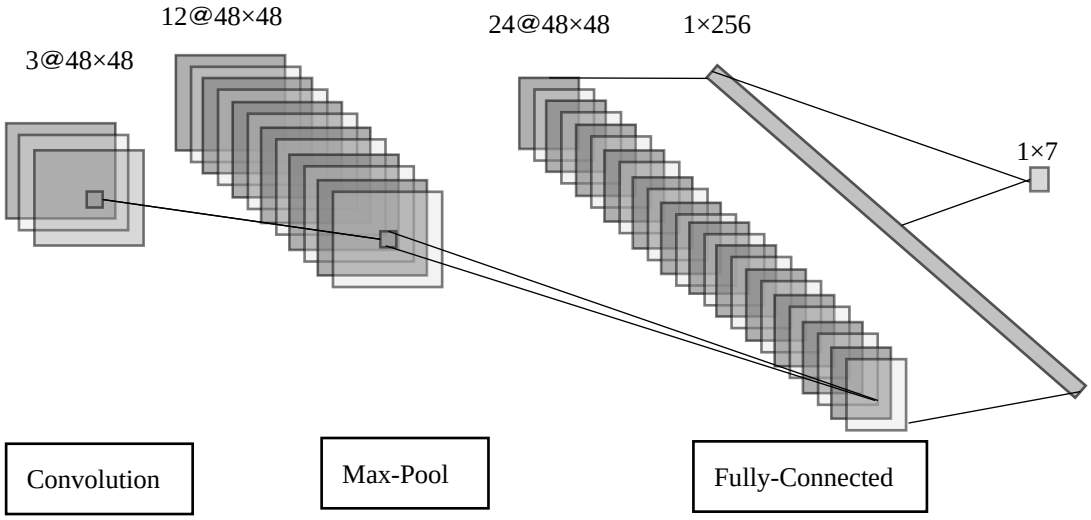

Figure 3. Schematic representation of the neural network

Fig. 3 depicts the configuration of the CNN network, which resembles the conventional AlexNet. The network comprises eight layers, where the initial five layers consist of a sequence of convolutional and pooling layers. The remaining three layers are responsible for classifying the CNN network based on the harmonic spectrum map, which is obtained by separating the raw music signal using HPSS. The input image size is standardized to 256 * 256, and the first convolution filter is applied. Following the flow of the deep network structure, 96 input images are examined, each with a size of 11 * 11 pixels, and 4 pixels are used for the first convolution layer's filtering. Subsequently, the output of the first convolution layer is fed into the maximum pool layer, which filters 96 cores with a size of 3 * 3. After response normalization, the second convolution layer is connected to its output, utilizing 256 cores with a size of 5 * 5 for connection and filtering. The third, fourth, and fifth convolution layers do not incorporate any pooling or normalization layers in between. The third convolution layer is connected to the first convolutional layers and has 384 cores of size 3 * 3. The fourth convolutional layer consists of 384 cores with a scale of 3 * 3, while the fifth convolutional layer comprises 256 cores with a specification of 3 * 3. By utilizing these five convolutional layers, 256 feature maps of size 6 * 6 are obtained. These feature maps are fed into three fully connected layers containing 4096, 1000, and 10 neurons, respectively. The output of the last fully connect layer represents the final recognition result. Additionally, Fig. 3 illustrates the underlying structure of CNNs, where the concealed layers consist of convolutional, pooling, and fully connected layers. It is crucial to acknowledge that there is no direct correlation between all neurons in the upper and lower layers of the CNN. CNNs possess distinct characteristics such as local connection, shared weights, and downsampling. These characteristics enable CNNs to extract local features from images effectively and exhibit strong resilience to image deformations. The convolutional layer, a pivotal component of the CNN, performs the crucial feature extraction task. Following convolution in the convolutional layer, the pooling layer reduces the dimensionality of the results from the upper layer, thereby decreasing the computational burden and the number of network training parameters. The fully connected layer is commonly positioned before the output layer and is crucial in categorizing multidimensional outcomes into one-dimensional data.

3.1.2.1 Convolutional layer

The convolutional layer is the central component of a CNN. Within this layer, the convolution core called the filter, is employed during the convolution process. With its specified size, this filter slides across the image horizontally and vertically, performing the convolution operation. As a result, a feature map is generated, representing the output of the convolutional layers, as depicted in Eq. (7).

$$feature_j^l = \phi\left(\sum_{i \in M_j} w_{ij}^l \otimes x_i^{l-1} + b_j^l\right) \quad (7)$$

In Eq. 7, $\otimes$ is the convolution operator, $feature_j^l$ is the output feature map, the J th is located in the low layer I, M_j is the set of output feature maps, w_{ij}^l is the convolution kernel of the J th feature map, namely the i th output data, w_{ij}^l is located in the first layer, where is the i th feature map, x_i^{l-1} is the deviation value of the j th feature map of the first layer, $\phi(\cdot)$ is the activation function. Commonly used activation functions are ReLU, sigmoid, tanh, etc.

The shallow convolutional layers [28] can generally obtain lower-level features like edges, boundaries, and lines. As the network hierarchy deepens, the deeper the convolutional layer can get, the more specific higher-level features.

This paper discusses the Sigmoid, Tanh, and ReLU activation functions, chosen for their relevance in different neural networks and their unique advantages. The Sigmoid and Tanh have been derived for RNNs because the output is mapped to some finite range, enabling a network to hold information over time, which is indispensable during sequential tasks in music recommendation. These functions, however, suffer from the gradient vanishing problem. On the other hand, ReLU remains a choice in CNNs due to its nature in the prevention of the gradient vanishing problem and is hence more appropriate for deep architecture. The linear growth of ReLU results in faster training and much more efficient convergence, particularly on tasks such as image and audio feature extraction using CNNs. By selecting these functions, the paper tries to emphasize the role of an activation function in optimization for network performance in music recommendation tasks.

3.1.2.2 Activation function

Utilizing an activation function in CNNs allows them to model nonlinearity. The network's capability is confined to representing linear mappings without an activation function. As a result, the representation of information in a multi-level CNN and a single or two-layer CNN becomes indiscernible. The following section elucidates the frequently employed activation functions.

The Sigmoid function is available across the entire domain, producing output values ranging from 0 to 1. Due to the Sigmoid function approaching 0 at both extremes, the range of function values changes very minimally. This can potentially result in the vanishing gradient problem, which hinders the backward propagation in DNNs. Eq. (8) depicts the Sigmoid function's mathematical representation.

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (8)$$

The Tanh function, the hyperbolic tangent function, is a mathematical function that exhibits odd symmetry. It produces output values within the range of -1 to 1 and possesses a gradient saturation effect. In contrast to the Sigmoid function, the Tanh function has an output mean of 0. Moreover, during neural network training, the Sigmoid function converges faster. The mathematical expression for the Tanh function is presented below.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (9)$$

The ReLU function is a segmented function, with 1 when x is greater than 0 and 0 when x is less than 0. The ReLU function effectively addresses the issue of gradient vanishing within the positive interval. In contrast to the Sigmoid and Tanh functions, the ReLU function exhibits linear growth. This characteristic enables faster calculations and significantly enhances the convergence rate compared to the sigmoid and Tanh functions. To tackle the gradient vanishing problem, many researchers have adopted the ReLU as the activation function. The mathematical expression for the ReLU function is presented below.

$$relu(x) = \begin{cases} x \; if x \geq 0 \\ 0 if \; x < 0 \end{cases} \quad (10)$$

3.1.2.3 Pooling layer

The pooling layer can reduce the dimension of each feature graph, although it does not guarantee that the most important information will not be lost. The pooling layer is commonly used at various stages

within the CNN. Typically, the pooling layer employs either the maximum or average pooling technique. To perform pooling on a 5 * 5 feature map, a pooling kernel of size 2 * 2 and a step size of 1 are defined. Similar to the convolution calculation procedure, the pooling procedure involves sliding the pooling kernel across the feature graph in a specified size, both horizontally and vertically. The pooling layer emphasizes the relative location between features rather than the precise location of the features themselves.

3.1.2.4 Full link layer

The fully connected layer transforms the high-dimensional feature maps generated through convolution and pooling into a lower-dimensional space. It accomplishes this by mapping the facial expression features learned by the CNN to the marker space of the dataset. The output of this layer is a set of n data points, each representing the probability of a specific type within the n categories. The highest probability among the n species is typically chosen as the final output.

3.2. Music Style Recognition Based on a Deep Learning CNN

While CNNs are widely recognized for their effectiveness in image recognition, their ability to detect spatial patterns also makes them valuable for analyzing audio data, such as music spectrograms. By treating audio spectrograms as images, CNNs can identify distinctive patterns associated with different music styles, enabling more precise genre classification. This capability allows CNNs to excel in music style recognition, a critical component of personalized music recommendation. In music software, the categorization of music styles holds significant value in reflecting a competitive advantage within the market. From the inception of music creation to today, no definitive demarcation exists between music styles and their variations. The evolution and integration of music styles persistently transpire, rendering music classification a formidable undertaking [29]. The pivotal task of classifying musical styles entails extracting style-related attributes from the music. Over the years, the CNN algorithms within the domain of deep learning have undergone continuous optimization to address the challenge of music classification. This research paper presents an exploratory music analysis method that relies on the CNN algorithm, yielding noteworthy outcomes. In the training and testing of this deep learning-based recommendation model, the dataset used was one comprising mostly English-language music tracks in the genres of pop, jazz, classical, rock, and electronic. Although this dataset provides a solid first testing ground, we do notice that it restricts the applicability of the model to a more global audience by excluding music from diverse linguistic and cultural contexts. We would improve our data set by including a wider range of cultural influences, such as traditional and modern Asian, African, and Latin American music, to allow the model to learn genre-specific idiosyncrasies of these traditions. In future iterations, we would like to collaborate with international music platforms and access public data sets from a variety of cultural backgrounds. Moreover, preprocessing involved conversion of audio tracks into representations of Mayer spectrum maps for more detail in harmonic and rhythmic features; this will enhance the model's ability to recognize complex audio patterns. This approach was chosen since it has been quite effective in music with regard to retaining temporal continuity, which is important for proper genre classification and recommendation.

The overall framework for music style identification based on CNNs is shown in Fig. 4. The initial layers of the CNN architecture serve as feature extractors, enabling the automatic acquisition of image features through supervised training. Subsequently, the recognition is classified using the softmax function in the final layer.In adapting the CNN structure for music style recognition, this study uses audio spectrograms as input, which are treated as if they were images to leverage the spatial pattern recognition capabilities that CNNs provide. The first few layers of the CNN learn low-level audio features, such as rhythm and pitch, with small convolutional filters that capture fine-grained details. As the network deepened, larger filters are used to capture higher-level, genre-specific features such as harmony and texture, which are very important to separate music styles. In order to extract features more effectively, HPSS is also used in preprocessing to separate the rhythm (percussive elements) from the melody (harmonic elements), which allow the CNN to be trained on different aspects of the music relevant for style classification. Additionally, dropout layers are used to prevent overfitting, ensuring the model could generalize effectively across diverse music genres.

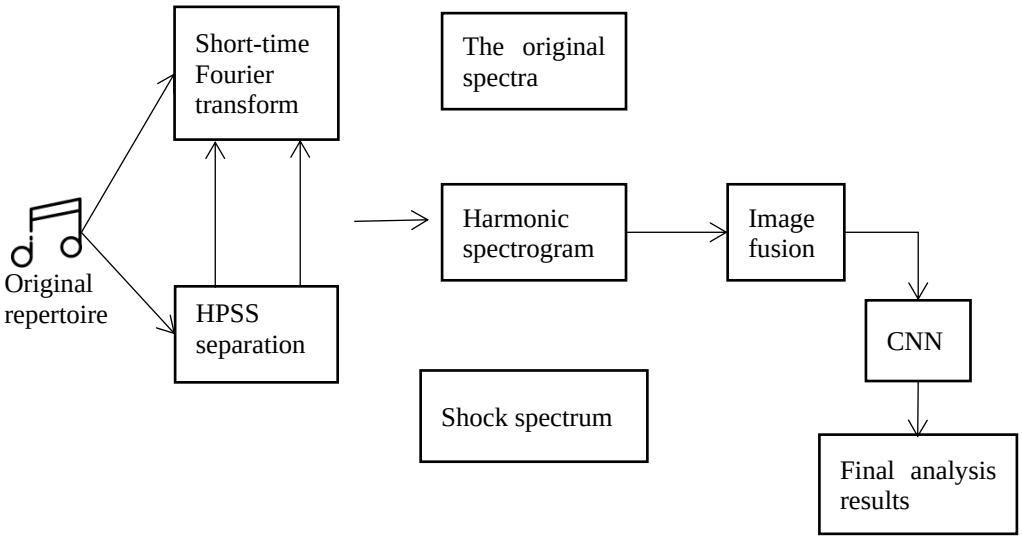

Figure 4. Schematic representation of the CNN

3.3. Recommendation Algorithm Based on Deep Learning

Traditional recommender systems generally rely on collaborative filtering—which recommends items based on similar users' preferences—or content-based filtering, which matches users to items based on item characteristics. However, there are some limitations to these methods: poor adaptability to dynamic changes in users' preferences, the cold start problem with new items, and an inability to capture the complex, nonlinear relations in user preferences.

These limitations are overcome by the deep-learning-based recommendation system introduced in this study, which uses a combination of CNNs for music style recognition and user interaction data to create a holistic view of user preferences. The CNNs extract features related to the genre and style directly from the spectrograms of the audio data, creating a rich feature representation for each track. Coupled with these features, the user behavioral data—hearing histories and skips, for example—provide a chance for the system to learn latent patterns in user preferences through neural network layers by modeling non-linear relationships.

This will result in more tailored and dynamic suggestions that are more aligned with users' tastes, which are evolutionary and reflective of interest in genre diversity, tackling a number of key challenges faced by traditional recommendation systems. Deep learning-powered recommendation systems typically utilize user and item data to train a deep learning model, enabling the extraction of latent representations for users and items. Subsequently, these hidden representations are leveraged to generate personalized user recommendation lists.

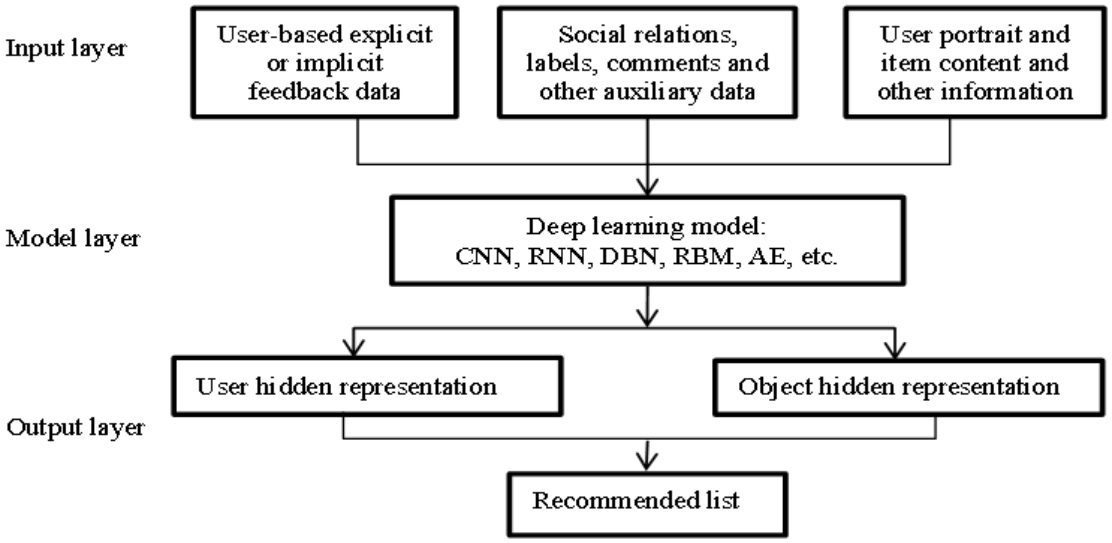

Figure 5. CNN structure process

Fig. 5 depicts the recommendation system's architecture based on deep learning. This system consists of three layers. The initial layer can receive explicit or implicit feedback data from users. This feedback data encompasses different aspects, such as user scores, browsing or clicking behavior, and user portrait and item content data, including preferences, age, picture, and audio content. The input data may also involve annotations, comments, and user-generated auxiliary data. CNNs, RNN's, autoencoders, and deep belief networks are a few of the deep learning models used by the model layer. These models play a pivotal role in acquiring the underlying representations of users and items. The primary function of the output layer is to generate a personalized list of recommended items for users. This is achieved by employing methods like Softmax classification and similarity calculation. The output layer produces accurate and tailored recommendations by integrating the acquired hidden representations of users and items.

4. Experimental Analysis

Within deep learning vision, many enterprises and organizations operating at national and international levels have introduced an extensive range of exceptional benchmark datasets specifically crafted for various application scenarios. These datasets, including MNIST, COCO, CIFAR, ImageNet, and Open Image, are widely utilized by researchers in this domain. The presence of these openly accessible datasets significantly contributes to the advancement and expansion of research and development in associated fields.

Nevertheless, in contrast to the vast availability of publicly accessible images or texts, the music information retrieval or recommendation domain faces a shortage of extensive, well-established, comprehensive, and user-friendly benchmark datasets. As a result, the advancement and implementation of models like DNNs, which typically require substantial data training, encounter certain obstacles within this field. Table 1 showcases a selection of the frequently accessible music datasets.

Table 1. Common publicly available music data sets

Data set	Sample number	Audio frequency	User records
Ballroom	698	Yes	No
GTZAN	1000	Yes	No
ISMIR2004	1458	Yes	No
MagnaTagAtune	25863	Yes	No
MSD	1000000	Interlinkage	Yes
AudioSet	2084320	Interlinkage	No
AcousticBrainz	2524739	No	No

The dataset, comprising 600 songs, is carefully curated to ensure a broad representation of styles, including pop, jazz, classical, rock, electronic, and folk. Selected from popular streaming services, these songs span both mainstream and niche genres, providing the model with a diverse range of rhythmic patterns, harmonic structures, and other stylistic elements. By including an equal number of songs from each genre, the dataset minimizes potential biases, fostering a balanced understanding of each genre's distinctive characteristics. This varied dataset enables the model to identify and learn distinguishing features across different musical styles, making it well-suited for training and testing in a deep learning-based recommendation system.

However, while publicly available datasets predominantly feature English songs, Chinese music data remains scarce despite its distinct qualities in melody, lyrics, and instrumentation. Chinese pop songs, for instance, often incorporate traditional instruments and unique singing techniques. To address this, 600 Chinese songs are randomly selected and added to the Cool Dog music client library. Over 60 days, the song-play record function in Cool Dog tracked the playback data of 12 users, generating a comprehensive dataset of user interactions. The resulting data, including playback counts and corresponding audio files, are preprocessed to capture the unique attributes of Chinese music for the recommendation model.

Several techniques are available for analyzing audio features, with the most commonly used ones being the sound spectrogram, Mayer spectrum, and Mayer inverted spectrum coefficient (MFCC). In recent years, the CNN has demonstrated remarkable capabilities in image processing, leading to the increased utilization of Mayer spectrum map features in DNN models for audio signal analysis. These features have gained more prominence than the MFCC. Consequently, this study aims to extract the log-Mayer spectrum map from the audio dataset and directly employ it as input for training subsequent network models. For instance, the

song "Happy Worship" exhibits a rhythmic pop style with a more hip-hop tempo, while "June Boat Song" is a pure piano composition accompanied by soothing tones. The Mayer spectrum features extracted from songs of the same style tend to be relatively similar, whereas those removed from different styles of songs exhibit significant differences. The reason the Mayer spectrum map is used in this research, and not MFCC, is that it retains both harmonic and rhythmic details critical for music analysis. Unlike MFCCs, which were originally tailored for speech processing and emphasize perceptible speech frequencies, the Mayer spectrum offers a more holistic representation of audio by collecting a wide range of frequencies. This makes it particularly well-suited for distinguishing music genres, where subtle harmonic and rhythmic nuances play a significant role. In addition, the Mayer spectrum provides temporal continuity, which is important to identify genre-specific patterns such as rhythm, beat, and tonal transitions. In this paper, through the proposed continuous structure of the Mayer spectrum, a CNN can effectively recognize sequential patterns of the music, which further improves its ability in style distinction. Furthermore, the Mayer spectrum has a spectrogram-like nature that is most compatible with CNN architectures optimized for image-like data. Such compatibility makes learning of hierarchical audio features—like complex textures and layers—much easier for the network compared to MFCCs, which essentially give a view of lower dimensionality and abstracted. This study, by utilizing the Mayer spectrum, allows for a richer and more detailed representation in features, which enables the model to perform robust music classification and recommendation based on diverse elements present in complex audio signals.

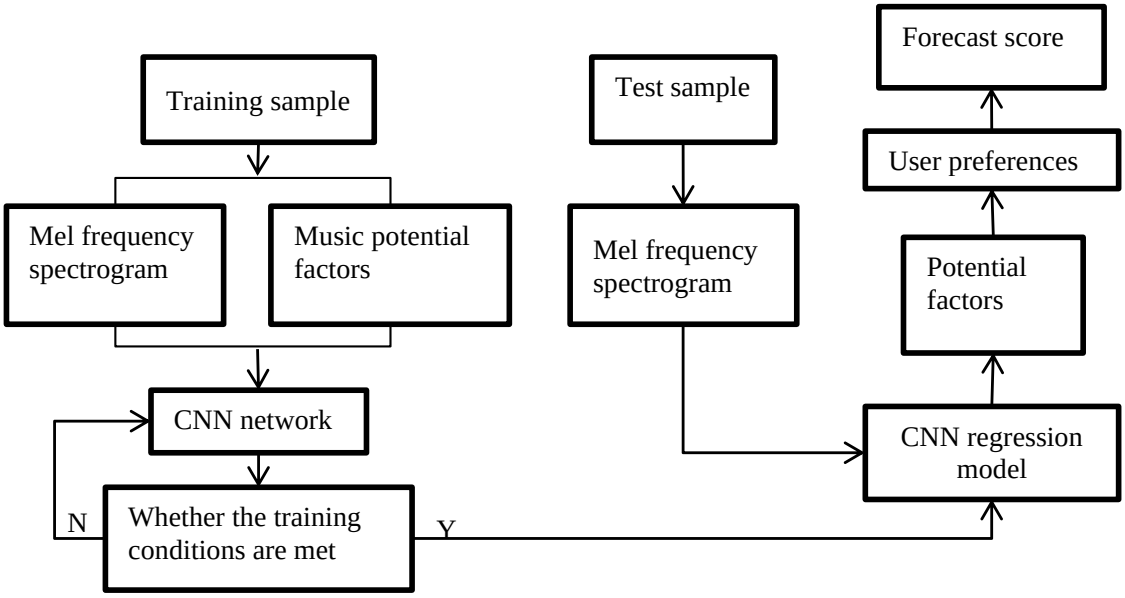

Figure 6. Flow chart of CNN network model training and testing

The HDF5 format is utilized in Keras to store the architecture and weights of a pre-trained CNN model, which can then be easily loaded for future use. Through the analysis of experimental results and comparisons, it becomes evident that deep learning music recommendation technology can effectively address the issue of a single characteristic representing a specific type of music information. This technology can accurately identify various music styles while retaining only the relevant information during the music recommendation process, enhancing the efficiency of music information prediction and making it more targeted. Furthermore, it performs better in classifying music information and exhibits enhanced feature learning and prediction capabilities compared to shallow networks. In the most comparative studies between deep learning models and shallow networks for music genre classification, deep learning models constantly outperform traditional shallow methods on key metrics. The accuracy rates of deep models, especially CNNs, usually reach about 90–92%, while the shallow models, like MLP and logistic regression, average around 70–78%. It highlights the superior ability of deep models in capturing complex audio features. Deep learning models usually have higher precision and recall; on many occasions, up to 90% and 85%, respectively, as opposed to the 70–75% range of the shallow networks. These suggest that the deep models are quite effective in identifying the right genres with fewer false positives. Similarly, in user engagement—measured by metrics like click-through rates (CTR) and time spent on recommended songs—

a 25–30% improvement can be observed with deep learning models, which results in better user retention and satisfaction due to the more accurate and personalized recommendations. Furthermore, deep learning models are better at dealing with cold start problems, with the latter achieving even up to 80–85% relevance in recommendation for new items, while shallow models usually plateau at around 60–65%. These results, as illustrated in research by Chen *et al*. [30] and Yin [31], show the vast improvements deep learning models provide in music recommendation systems over traditional shallow networks.

5. Conclusion

With the continuous development of the Internet and technology, music recommendation has become one of the most important parts in improving the market competitiveness and user experience for major music platforms. Obviously, the effectiveness of a recommendation system is critical to user satisfaction, which has driven increased research interest in music recognition and recommendation techniques. In recent years, deep learning has been the hot topic in the field of artificial intelligence and has obvious advantages in speech recognition, audio processing, and many other fields due to its solid learning capabilities.

Based on deep learning principles, this paper proposes a method for music recognition and recommendation that relies on deep confidence networks to overcome the drawbacks of the traditional recommendation system. Via multi-feature collection, this approach enhances music style recognition and diversifies recommendation, thus allowing more personalized and varied suggestions of music. The results show that such an approach improves the multi-feature recognition of music styles to create a more adaptive recommendation system capable of serving users with changing preferences better.

One limitation of the current study is its reliance on an English-language dataset, which limits the model's applicability in multicultural contexts. Music preferences are strongly determined by culture, and a dataset that covers music with diverse linguistic and regional backgrounds—like traditional and popular music from Asia, Africa, and Latin America—could make the model more genre-specific and increase its applicability to an international audience. Moreover, the complexity in the CNN model makes it a scalability challenge, especially for the smaller platform with less computation resources. Generally, deep learning models, including CNNs, require high processing power to be executed, which might become a barrier to their adoption on platforms without access to high-performance computing infrastructure. Future research would therefore focus on these limitations in an attempt to improve dataset diversity and model optimization techniques, which include pruning, quantization, and lightweight architectures such as MobileNet. These strategies are used to explore other tasks across different domains.

Future work will focus on refining prediction accuracy and expanding the system's classification capabilities. Potential enhancements include the use of deeper architectures, such as Transformer models, which could better capture intricate temporal and spectral patterns in audio data. Additionally, incorporating advanced feature-engineering techniques, such as tempo and rhythm analysis, and addressing data sparsity through matrix completion algorithms, may improve generalization across diverse genres and user behaviors. These steps aim to elevate the system's adaptability, setting a robust foundation for future music recommendation applications. Also, Future work will involve expanding the dataset with a focus on music from an increasingly diverse number of cultures and languages represented. Adding music from regions like Asia, Africa, and Latin America, the model could grasp even more of the variety of musical styles, hence making it more applicable globally and better at serving diverse user preferences. Further, it is targeted optimization techniques for models in order to meet scalability issues, especially in the case of low-computation platforms. Model pruning and quantization techniques are evaluated, and lightweight architectures like MobileNet and EfficientNet are experimented with to reduce the computational demands of the model. Second in line will be experiments in real-world settings, including mobile and low-power environments, to test whether it is practical and adaptive to different deployment scenarios. Such developments would lead to the creation of a more inclusive, resource-efficient recommendation system suitable for a larger audience and different applications.

Competing of Interests

The authors declare no competing of interests.

References

[1] Robert Prey, Marc Esteve Del Valle and Leslie Zwerwer, 'Platform pop: disentangling Spotify's intermediary role in the music industry', *Information, Communication & Society*, Vol. 25, No. 1, pp. 74–92, 22nd May 2020, ISSN: 1468-4462, DOI: 10.1080/1369118X.2020.1761859, Available: https://www.tandfonline.com/doi/full/10.1080/1369118X.2020.1761859.

[2] Zhen Troy Chen, 'Flying with two wings or coming of age of copyrightisation? A historical and socio-legal analysis of copyright and business model developments in the Chinese music industry', *Global Media and China*, Vol. 6, No. 2, pp. 191–206, 1st March, 2021, Online ISSN: 2059-4372, DOI: 10.1177/2059436421998466, Available: https://journals.sagepub.com/doi/full/10.1177/2059436421998466.

[3] Malin Song, Chenbin Zheng and Jiangquan Wang, 'The role of digital economy in China's sustainable development in a post-pandemic environment', *Journal of Enterprise Information Management*, Vol. 35, No. 1, pp. 58–77, 18th February 2022, ISSN: 1741-0398, DOI: 10.1108/JEIM-03-2021-0153, Available: https://www.emerald.com/insight/content/doi/10.1108/jeim-03-2021-0153/full/html.

[4] Bindu Balagopal and Chacko Jose P, 'Innovations in digital technology and creative destruction in the music industry', *International Journal of Indian Culture and Business Management*, Vol. 24, No. 3, pp. 303–318, 10th December 2021, ISSN: 1753-0814, DOI: 10.1504/IJICBM.2021.119737, Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJICBM.2021.119737.

[5] Christopher Buccafusco and Kristelia García, 'Pay-to-Playlist: The Commerce of Music Streaming', *UC Irvine Law Review*, Vol. 12, p. 805, 11th March 2021, ISSN: 2327-4514, DOI: 10.2139/ssrn.3793043, Available: https://heinonline.org/HOL/LandingPage?handle=hein.journals/ucirvlre12&div=25&id=&page=.

[6] Mingli Shang and Hui Sun, 'Study on the New Models of Music Industry in the Era of AI and Blockchain', in *Proceedings of the 2020 3rd International Conference on Smart BlockChain (SmartBlock)*, 23-25 October 2020, Zhengzhou, China, pp. 63–68, Electronic ISBN: 978-1-6654-4073-8, DOI: 10.1109/SmartBlock52591.2020.00019, Available: https://ieeexplore.ieee.org/abstract/document/9415657.

[7] Rahib Abiyev, John Bush Idoko and Murat Arslan, 'Reconstruction of convolutional neural network for sign language recognition', in *Proceedings of the 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, IEEE, 12-13 August 2020, Istanbul, Turkey, pp. 1–5, Electronic ISBN: 978-1-7281-7116-6, DOI: 10.1109/ICECCE49384.2020.9179356, Available: https://ieeexplore.ieee.org/abstract/document/9179356.

[8] Robi Polikar, Lalita Upda, Satish S Upda and Vasant Honavar, 'Learn++: An incremental learning algorithm for supervised neural networks', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 31, No. 4, pp. 497–508, 30th November 2001, Electronic ISSN**:** 1558-2442, DOI: 10.1109/5326.983933, Available: https://ieeexplore.ieee.org/abstract/document/983933.

[9] David Goldberg, David Nichols, Brian M Oki and Douglas Terry, 'Using collaborative filtering to weave an information tapestry', *Communications of the ACM*, Vol. 35, No. 12, pp. 61–70, 1st December 1992, ISSN: 0001-0782, DOI: 10.1145/138859.138867, Available: https://dl.acm.org/doi/abs/10.1145/138859.138867.

[10] Vasil Vatakov, Evelina Pencheva and Emilia Dimitrova, 'Recent advances in artificial intelligence for improving railway operations', in *Proceedings of the 2022 30th National Conference with International Participation (TELECOM)*, 27-28 October 2022, Sofia, Bulgaria, pp. 1–4, Electronic ISBN: 978-1-6654-8212-7, DOI: 10.1109/TELECOM56127.2022.10017265, Available: https://ieeexplore.ieee.org/abstract/document/10017265.

[11] Halie Rando, Nils Wellhausen, Soumita Ghosh, Alexandra Lee, Anna Dattoli *et al.*, 'Identification and Development of Therapeutics for COVID-19', *Msystems*, Vol. 6, No. 6, pp. e00233-21, 2nd November 2021, ISSN: 2379-5077, DOI: 10.1128/mSystems.00233-21, Available: https://journals.asm.org/doi/full/10.1128/msystems.00233-21.

[12] Hongyang Shi, Yu Mei, Ian González-Afanador, Claudia Chen, Scott Miehls *et al.*, 'Automated Soft Pressure Sensor Array-Based Sea Lamprey Detection Using Machine Learning', *IEEE Sensors Journal*, Vol. 23, No. 7, pp. 7546–7557, 2nd March 2023, Electronic ISSN: 1558-1748, DOI: 10.1109/JSEN.2023.3249625, Available: https://ieeexplore.ieee.org/abstract/document/10058136.

[13] Youlan Tao, Hui Wen and Shuhuai Wang, 'Translation teaching research in the Chinese mainland (1978–2018): Theory, method and development', in Key Issues in Translation Studies in China: Reflections and New Insights, Singapore: Springer Nature, 27th June 2020, Ch. 3, pp. 47–76, Print ISBN: 978-981-15-5864-1, Online ISBN: 978-981-15-5865-8, DOI: 10.1007/978-981-15-5865-8_3, Available: https://link.springer.com/chapter/10.1007/978-981-15-5865-8_3.

[14] Ya Zhou, Yao Hu, Liquan Dong, Yuejin Zhao, Yong Song *et al.*, 'Optoelectronic instrument experiments course: A trial of project-based learning', in *Proceedings of the 2012 7th International Conference on Computer Science & Education (ICCSE)*, IEEE, 14-17 July 2012, Melbourne, VIC, Australia, pp. 1375–1379, Electronic ISBN: 978-1-4673-0242-5, DOI: 10.1109/ICCSE.2012.6295319, Available: https://ieeexplore.ieee.org/abstract/document/6295319.

[15] Dongliang Fan, Xiaoyun Su, Bo Weng, Tianshu Wang and Feiyun Yang, 'Research progress on remote sensing classification methods for farmland vegetation', *AgriEngineering*, Vol. 3, No. 4, pp. 971–989, 8th December 2021, ISSN: 2624-7402, DOI: 10.3390/agriengineering3040061, Available: https://www.mdpi.com/2624-7402/3/4/61.

[16] Zhuo-Hang Lv, Han-Bing Yan and Rui Mei, 'Automatic and accurate detection of webshell based on convolutional neural network', in *Proceedings of the China Cyber Security Annual Conference*, 14th – 16th August 2018, Beijing, China, Vol. 970, pp. 73–85, ISBN: 978-981-13-6621-5, DOI: 10.1007/978-981-13-6621-5_6, Available: https://link.springer.com/chapter/10.1007/978-981-13-6621-5_6.

[17] Zhongyuan Jia, 'Designing an Intelligent Teaching System of Chinese as a Foreign Language under the Internet Background', *Scientific Programming*, Vol. 2022, pp. 1-10, 30th July 2022, ISSN: 1058-9244, DOI: 10.1155/2022/3610081, Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/3610081.

[18] Hojjat Adeli, *Expert systems in construction and structural engineering*, 1st ed. Boca Raton, FL, USA: CRC Press, 3rd March 1988, ISBN: 9780429083327, DOI: 10.1201/9781482289008, Available: https://www.taylorfrancis.com/books/mono/10.1201/9781482289008/expert-systems-construction-structural-engineering-adeli.

[19] Siyuan Liu, *Xin Fengxia and the Transformation of China's Ping Opera*, 1st ed. London, UK: Cambridge University Press, 26th August 2022, Online ISBN: 9781009083508, DOI: 10.1017/9781009083508, Available: https://www.cambridge.org/core/elements/abs/xin-fengxia-and-the-transformation-of-chinas-ping-opera/A4DEF22128DCC19AA865029131353AB0.

[20] Yoon Kim, Yacine Jernite, David Sontag and Alexander Rush, 'Character-aware neural language models', in *Proceedings of the AAAI Conference on Artificial Intelligence*, 5th March 2016, Phoenix, Arizona, USA, ISSN: 2374-3468, DOI: 10.1609/aaai.v30i1.10362, Available: https://ojs.aaai.org/index.php/AAAI/article/view/10362.

[21] Yu Zhang and Binglong Li, 'Malicious code detection based on code semantic features', *IEEE Access*, Vol. 8, pp. 176728–176737, 23rd September 2020, Electronic ISSN: 2169-3536, DOI: 10.1109/ACCESS.2020.3026052, Available: https://ieeexplore.ieee.org/abstract/document/9204732.

[22] Rui Hou, 'Music content personalized recommendation system based on a convolutional neural network', *Soft Computing*, Vol. 28, No. 2, pp. 1785–1802, January 2024, ISSN: 1433-7479, DOI: 10.1007/s00500-023-09457-2, Available: https://link.springer.com/article/10.1007/s00500-023-09457-2.

[23] Xinglin Wen, 'Using deep learning approach and IoT architecture to build the intelligent music recommendation system', *Soft Computing*, Vol. 25, No. 4, pp. 3087–3096, February 2021, ISSN: 1433-7479, DOI: 10.1007/s00500-020-05364-y, Available: https://link.springer.com/article/10.1007/s00500-020-05364-y.

[24] Yezi Zhang, 'Music recommendation system and recommendation model based on convolutional neural network', *Mobile Information Systems*, Vol. 2022, No. 1, p. 3387598, 12th May 2022, ISSN: 1875-905X, DOI: 10.1155/2022/3387598, Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/3387598.

[25] John J Hopfield, 'Neural networks and physical systems with emergent collective computational abilities.', *Proceedings of the National Academy of Sciences*, Vol. 79, No. 8, pp. 2554–2558, 15th April 1982, ISSN: 1091-6490, DOI: 10.1073/pnas.79.8.2554, Available: https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

[26] Dylan Molho, Jiayuan Ding, Wenzhuo Tang, Zhaoheng Li, Hongzhi Wen *et al.*, 'Deep learning in single-cell analysis', *ACM Transactions on Intelligent Systems and Technology*, Vol. 15, No. 3, pp. 1-62, 29th March 2022, ISSN: 2157-6904, DOI: 10.1145/3641284, Available: https://dl.acm.org/doi/abs/10.1145/3641284.

[27] Hongjie Yi, Guangrong Ji and Haiyong Zheng, 'Optimal Parameters of Bp Network for Character Recognition', in *Proceedings of the 2012 International Conference on Industrial Control and Electronics Engineering*, 23-25 August 2012, Xi'an, China, pp. 1757–1760, ISBN: 978-1-4673-1450-3, DOI: 10.1109/ICICEE.2012.465, Available: https://ieeexplore.ieee.org/abstract/document/6322755.

[28] Mohammad Mahdi Bejani and Mehdi Ghatee, 'A systematic review on overfitting control in shallow and deep neural networks', *Artificial Intelligence Review*, Vol. 54, No. 8, pp. 6391–6438, 3rd March 2021, ISSN: 0269-2821, DOI: 10.1007/s10462-021-09975-1, Available: https://link.springer.com/article/10.1007/s10462-021-09975-1.

[29] Che Liu, Qian Ma, Zhang J. Luo, Qiao R. Hong, Qiang Xiao *et al.*, 'A programmable diffractive deep neural network based on a digital-coding metasurface array', *Nature Electronics*, Vol. 5, No. 2, pp. 113–122, February 2022, ISSN: 2520-1131, DOI: 10.1038/s41928-022-00719-9, Available: https://www.nature.com/articles/s41928-022-00719-9.

[30] Ke Chen, Beici Liang, Xiaoshuan Ma and Minwei Gu, 'Learning audio embeddings with user listening data for content-based music recommendation', in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 06-11 June 2021, Toronto, ON, Canada, pp. 3015–3019, Electronic ISSN: 2379-190X, DOI: 10.1109/ICASSP39728.2021.9414458, Available: https://ieeexplore.ieee.org/abstract/document/9414458.

[31] Tingrong Yin, 'Music Track Recommendation Using Deep-CNN and Mel Spectrograms', *Mobile Networks and Applications*, Vol. 28, pp. 1–8, December 2023, ISSN: 1383-469X, DOI: 10.1007/s11036-023-02170-2, Available: https://link.springer.com/article/10.1007/s11036-023-02170-2.