Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **4**, November **2025**

A Study on Music Genre Classification Model Based on Convolutional Neural Network and Transfer Learning

Yi-Ran Zhang*

Department of Music
Nanchang Institute of Technology
Nanchang 330044, P. R. China
sax1135@163.com

Yu-Han Wang

International College
Krirk University
Bangkok 10220, Thailand
zitong0825@gmail.com

*Corresponding author: Yi-Ran Zhang
Received July 15, 2024, revised December 9, 2024, accepted July 4, 2025.

Abstract. *In the wave of rapid development of artificial intelligence technology, music genre classification, as a key branch in the field of music information retrieval, is gradually becoming a research hotspot. This paper proposes an innovative automatic music genre classification model based on Convolutional Neural Network (CNN) and migration learning, aiming at realizing efficient automatic feature extraction and accurate classification of music signals through deep learning technology. The article first provides a comprehensive review of the research background, current status and its challenges in music genre classification, clearly points out the limitations of traditional methods, and emphasizes the potential of deep learning in improving the performance of music genre classification. Subsequently, the paper details the proposed model architecture, covering all aspects from data preprocessing to feature extraction, model training, and classifier design. In particular, in the feature extraction stage, the automatic learning capability of CNNs is fully utilized to capture the complex time-frequency features of music signals. To enhance the generalization of the model, this paper employs a transfer learning strategy to adapt a pre-trained CNN model to the music genre classification task. By fine-tuning on a wide range of datasets, the model not only quickly adapts to the new classification task, but also significantly reduces the training cost. In the experimental part, the proposed model is comprehensively evaluated on a publicly available music genre dataset in this paper. The evaluation results show that the proposed model exhibits significant advantages in terms of classification accuracy, robustness, and computational efficiency compared with the existing techniques. In addition, this paper also provides an in-depth analysis of the interpretability of the model and explores the specific impact of different hierarchical features on classification decisions. The article concludes with a summary of the research results and an outlook on future research directions. Given the complexity of the music genre classification task, future research will focus on further optimization of the model, fusion processing of multimodal data, and exploration of cross-domain applications.*
Keywords: convolutional neural network; transfer learning; music genre classification; deep learning; automatic feature extraction

1. **Introduction.** Music, as a bright pearl in the treasure house of human culture, the diversity and richness of its genres have always been the focus of musicology and the music industry. Driven by the wave of digitization, music data has shown explosive growth, which makes the automatic identification and classification technology of music genres particularly important [1]. It is not only the core technology in the fields of music information retrieval, personalized recommendation systems and music education, but also the key to promote the development of the music industry. However, traditional music genre classification methods rely on hand-selected features, and these methods are often limited by the subjectivity of feature selection, making it difficult to capture the nuances and rapid changes in music styles. Therefore, the development of a more efficient and accurate automated music genre classification method not only has far-reaching academic significance, but also possesses broad application prospects [2].

In recent years, deep learning techniques have made breakthroughs in a number of fields, including image, speech and text processing. In particular, Convolutional Neural Networks (CNNs), with their superior feature extraction capabilities, provide a new perspective for automatic music genre classification [3]. CNNs are able to automatically mine the deep features of the data without relying on complex manual feature engineering, which opens up a new technical path for music genre classification [4]. Nevertheless, the high dimensionality and complexity of music data pose challenges to the direct application of CNNs, including the difficulty of model training and high training costs. Migration learning, an innovative machine learning strategy, allows models to migrate knowledge learned in one domain to another, thus significantly improving the generalization ability and learning efficiency of the models. In the context of music genre classification, by migrating pre-trained CNN models, we are able to achieve rapid adaptation and learning for new tasks while maintaining low training costs [5].

In this study, we propose a music genre classification model based on convolutional neural networks and transfer learning (CNN-TL). Through an in-depth analysis of the characteristics of music signals, we design a set of optimized CNN architectures to achieve efficient automatic feature extraction of music signals. In addition, through the transfer learning technique, we adapt the pre-trained CNN model to the music genre classification task, and further improve the classification accuracy and generalization ability of the model by fine-tuning it on specific datasets. The results of this study not only provide an innovative technical solution in the field of music genre classification, but also inject new vigor into the development of music information retrieval technology.

1.1. **Related work.** As deep learning technology continues to advance at a swift pace, CNNs have emerged as a preferred solution for intricate pattern recognition challenges, including the categorization of music genres. The remarkable proficiency of CNNs in domains like image recognition and natural language processing offers innovative avenues for the automated classification of musical styles [6]. In research on music genre classification, scholars have begun to explore how to utilize the powerful feature extraction capabilities of CNNs, combined with the uniqueness of music data, to achieve more accurate classification. For example, Johnson and Zhang [7] introduced bag-of-words model transformation in CNN to enhance the expressive ability of text features, an innovation that achieved significant accuracy improvement in text classification tasks. Liao et al. [8] demonstrated, through comparative analysis, that CNN-based methods are more effective in the task of sentiment categorization of Twitter data compared to traditional machine learning methods such as Support Vector Machine (SVM) and naive Bayesian algorithms, with higher accuracy. In addition, Jamatia et al. [9] utilized a neural network model to process hate speech text on Twitter, and experimentally verified the superiority

of the word2vec embedding-based model in the text categorization task. Pereira et al. [10] employed a multiscale information processing strategy to efficiently perform sentiment analysis of short texts, which provided a new music genre classification thinking perspective.

However, traditional machine learning methods are usually based on the assumption that the training and test data have the same feature space and consistent distribution, which is often difficult to fulfill in real-world applications. To overcome this challenge, this study introduces the idea of transfer learning and draws on the innovative transfer learning method proposed by Zhang et al. [11]. It effectively improves the generalization ability of classification algorithms by identifying variable and invariant features across different datasets. This study further explores the successful case of applying migration learning in visual recognition tasks by Luo et al. [12]. The image feature representations obtained from CNN training on large-scale labeled datasets are migrated to other tasks, and good results can be achieved even with limited training data. Liu et al. [13] applied CNN-based migration learning methods to the problem of text recognition of historical documents. The problem of insufficient labeled training samples was solved. Raffel et al. [14] trained on the ImageNet dataset by migration learning algorithm, which in turn achieved performance improvement in image detection and classification tasks.

These research results not only demonstrate the potential of CNN combined with transfer learning in improving the accuracy of classification tasks, but also prove its effectiveness in improving the generalization ability and utility of the model when the data distribution is inconsistent or the training samples are scarce. By constructing a music genre classification model based on CNN and migration learning, this study provides a new technical solution in the field of music information retrieval, and lays the foundation for the future application of deep learning techniques in a wider range of music scenarios.

1.2. **Motivation and contribution.** Music genre classification, as a cornerstone in the field of music information retrieval, plays an indispensable role in building an efficient intelligent music recommendation system, significantly enhancing users' music experience, and driving the innovation and growth of the music industry. Although previous research has achieved preliminary results in this area, existing techniques still face many challenges in dealing with the growing large-scale and diverse music datasets. In particular, existing methods show significant limitations in terms of the accurate extraction of musical features and the ability of models to generalize to new data. To address these challenges, this study proposes an innovative solution that utilizes deep learning techniques, especially convolutional neural networks (CNNs) and transfer learning, to break through the bottleneck of existing techniques. With this approach, we are not only able to automatically extract deep features of music signals, but also significantly improve the model's adaptability and classification accuracy for different music datasets. The goal of this research is to realize the comprehensive improvement of music genre classification performance through these advanced technological means, and to bring new breakthroughs in the field of music information retrieval.

The contributions of the paper are mainly in the following areas:

1. We propose a CNN-based music feature extraction method that automatically learns a deep representation of music signals without relying on traditional manual feature engineering. Compared with existing methods, our method is able to capture the time-frequency characteristics and rhythmic patterns of music signals more accurately, providing richer and more effective features for music genre classification.
2. We innovatively apply the transfer learning technique to the music genre classification task, and significantly improve the model's adaptability to new datasets by

fine-tuning on a large-scale pre-trained model. This approach not only reduces the reliance on large amounts of labeled data, but also reduces the cost and time of model training.

3. By systematically analyzing and adjusting the degree of transfer learning, our model demonstrated excellent classification performance and good generalization ability on publicly available music genre datasets, validating the practicality and flexibility of the model.

2. Relevant theoretical analysis.

2.1. Convolutional neural networks.

Convolutional neural networks are developed from the multilayer perceptual machine model [15], LeNet-5. In 2012, the ImageNet competition was revolutionized by AlexNet's triumph in the classification challenge, sparking a surge in the adoption of CNNs across various real-world applications [16]. A CNN, characterized by its deep, feed-forward architecture, stands as a quintessential deep learning algorithm. It has since gained extensive application in domains such as image processing and natural language processing. The traditional architecture of a CNN typically encompasses several key components: an input layer to receive raw data, convolutional layers for feature extraction, pooling layers to reduce spatial dimensions, fully connected layers for high-level processing, activation functions to introduce non-linearity, and an output layer to produce the final predictions [17].

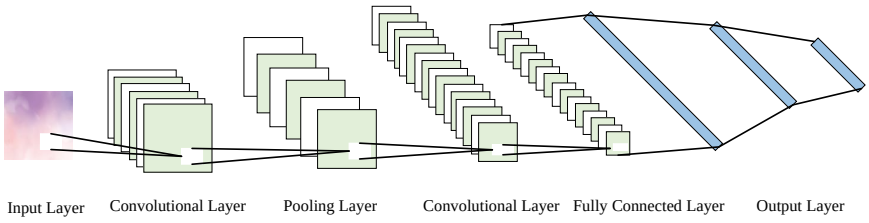

Figure 1. Convolutional neural network model.

A convolutional layer is the most basic structural unit in a convolutional neural network model, which utilizes multiple convolutional kernels to do convolutional processing of the input data and extract image features. Shallow convolutional layers usually extract only detailed features such as lines, edges, corners, etc., while deeper convolutional layers iterate and integrate these shallow features to obtain abstract features. The mathematical expression for convolutional layer is shown in Equation (1):

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * \omega_{ij}^l + b_j^l\right) \tag{1}$$

where ω_{ij}^l and b_j^l denote the weight and bias corresponding to the convolutional filter at position (i, j), respectively; x_i^{l-1} denotes the feature mapping of the previous layer; x_j^l denotes the feature mapping of the current layer; $f(\cdot)$ denotes the activation function; and M_j denotes the set of feature mappings.

Typically, an activation function is added between the convolutional and pooling layers for nonlinear fitting operations on the input data to enhance the approximation ability and nonlinearity of the network. Typical activation functions are sigmoid, tanh, ReLU, etc. The output value of sigmoid is in the range of [0,1], which is equivalent to normalizing the output value of each neuron. The sigmoid is calculated as shown in Equation (2):

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

The output value of tanh is in the range of [-1,1]; specifically, the output value is 1 when the positive number is larger and -1 when the negative number is larger. The calculation of tanh is shown in Equation (3):

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3}$$

When the input value is less than 0, the output value of ReLU is 0 and the derivative value is also 0. This will lead to the neuron not being able to perform parameter update and the phenomenon of gradient vanishing. The calculation of ReLU is shown in Equation (4):

$$ReLU(x) = \max(0, x) \tag{4}$$

Currently, the average pooling operation and the maximum pooling operation are the two pooling operations frequently used in constructing network models. According to Equation (5), the average pooling operation takes the average value corresponding to the pooled region; the maximum pooling operation takes the maximum value corresponding to the pooled region, and its mathematical expression is shown in Equation (6):

$$pooling_{u,v}^{max} = \frac{1}{|\Omega_{u,v}|} \sum_{i,j \in \Omega_{u,v}} a_{i,j}, \tag{5}$$

$$pooling_{u,v}^{average} = \max_{i,j \in \Omega_{u,v}} a_{i,j} \tag{6}$$

where $a_{i,j}$ is the activation value of the pooled region; i, j is the index representation; and $\Omega_{u,v}$ is the corresponding pooled region on the feature map.

Each neuron within the fully connected layer is interconnected with those of the preceding layer, establishing a comprehensive network of connections. Post the pooling stage, the resultant feature map from the prior layer undergoes a transformation into a one-dimensional array. This vectorized form is then conveyed to the fully connected layer, which is tasked with distilling higher-level, abstract features essential for the conclusive classification phase. Typically, the fully connected layer is positioned as the final component of the network architecture to augment its overall efficacy. In the contemporary landscape of convolutional neural networks, the softmax function stands as a fundamental and prevalent classification tool. When addressing multi-class classification challenges, the softmax function forecasts the conditional likelihood of an input instance x pertaining to a specific class c. This probability is articulated through the formulation presented in Equation (7):

$$p(y = c \mid x) = \frac{e^{w_c^T x}}{\sum_{j=1}^{C} e^{w_j^T x}} \tag{7}$$

where C denotes the total number of categories; w_c represents the weight vector for category c; and p represents the conditional probability of the category belonging to category c.

At the culmination of a CNN lies the output layer, which is pivotal in the realm of classification tasks. The neuron count within this terminal layer is aligned with the variety of classification classes, ensuring a direct correlation. Subsequently, a softmax activation function is applied to this layer, serving to ascertain the likelihood of the input image

being assigned to each of the potential categories, thereby generating a comprehensive probability distribution. After the softmax activation function processing maps the network model's score for the input data to the (0,1) interval, the output of the softmax activation function is the classification probability of the network model for the input samples in each category. The expression of the softmax activation function is shown in Equation (8):

$$a_i = \frac{e^{z_i}}{\sum_{k=1}^{m} e^{z_k}} \tag{8}$$

where z_i is the score of the network model for category i and a_i is the predicted probability value of the input on category i. The maximum probability value of a_i determines the category of the prediction result, and the sum of all a_i is 1. The network model can calculate the loss value for this training by using the probability distribution of the output and the true label of the input samples during the training process.

2.2. **Transfer learning.** Transfer learning was incorporated into machine learning in the 1990s and is a new paradigm in machine learning [18]. First, we will delve into the basic concepts of domain and task in order to better understand the relationship between them. The source domain is a dataset that has already been learned, the source task is a task that has already been completed, and the target task is an upcoming task. The target domain is a dataset that is about to start learning. These two concepts will be elaborated in depth next [19].

Definition 1: A domain D consists of two parts, the feature space X and the edge probability density $P(X)$, where $x \in X$.

Definition 2: Task T is constituted by two fundamental elements: the set of labels, denoted as y, and a decision-making process encapsulated by the function f. This is formally expressed as $T = \{y, f\}$. The function f is derived through the learning process from the dataset at hand, rather than being pre-defined. Within the scope of this study, the function f symbolizes the classification mechanism specific to categorizing music genres.

Definition 3: Providing a clear learning scope, including the source domain D_s, the target domain D_t, and the tasks T_s and T_t, as well as a certain learning style for better understanding and mastering knowledge. Hence, the concept of transfer learning can be interpreted as the application of expertise acquired from the training process within a source domain and task to facilitate the learning of the predictive function within the target domain. This is achieved by leveraging the insights from the source domain and its associated task conjunction with the target domain and its specific task.

$$T(t) = f(D_t) \tag{9}$$

where $D_s \neq D_t$, $T_s \neq T_t$. Figure 2 shows the schematic diagram of transfer learning.

2.3. **ResNet model.** The source network migrated in this study is the Deep Residual Network (ResNet). A short-circuit connected network structure was used, as shown in Figure 3.

Figure 3 shows the core network block that constitutes the ResNet network, where the weight layer represents the feature extraction layer such as the convolutional layer in the network; ReLU is the activation function used by the network; x and $F(x)$ are the features input to the block and the features obtained after the two-layer operation. The core of the block's design is the right-hand side of the short-connection structure, so that the feature x before input to the network and the feature $F(x)$ after operation are retained at the same time [20]. It is guaranteed that after the operation of this network block, at

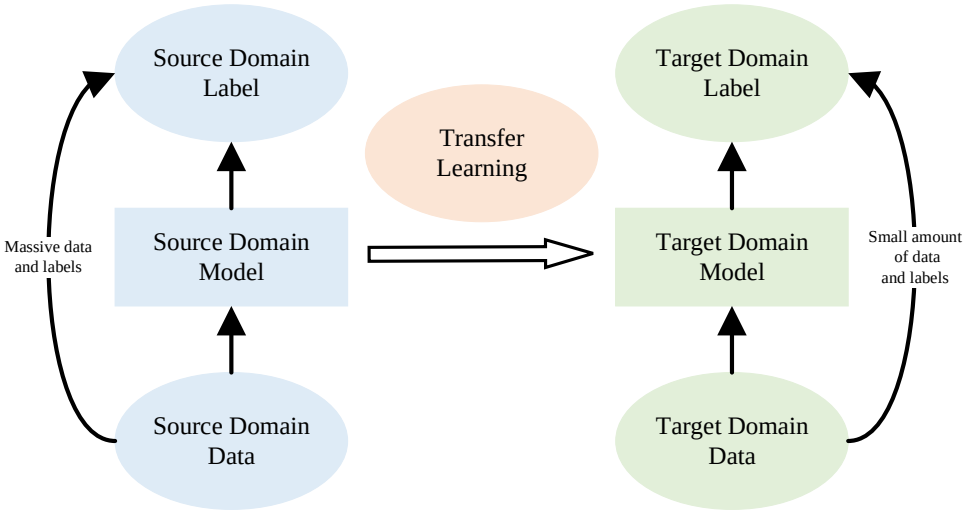

Figure 2. Schematic diagram of Transfer Learning.

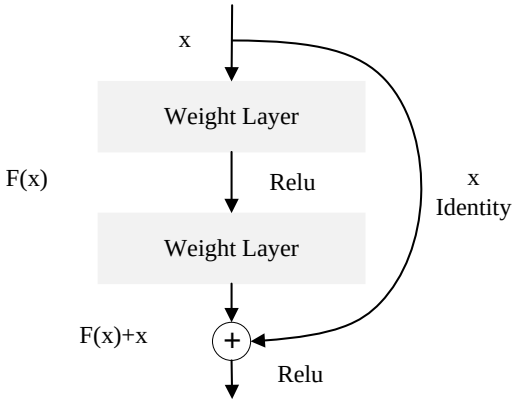

Figure 3. The basic composition of the residual network.

least the existing feature extraction results will not be lost and the degradation problem is avoided. Where the residual block can be shown as follows:

$$y_l = h(x_l) + F(x_l, \omega_{x_l}) \tag{10}$$

$$x_{l+1} = f(y_l) \tag{11}$$

where x_l and x_{l+1} represent, respectively, the input and output of the lth residual unit, where each residual unit usually contains a multilayer convolutional structure. f is the residual function, which represents the residuals learned by the network. In addition, $h(x_l) = x_l$ denotes the constant mapping and f denotes the ReLU activation function. Incorporating a bypass connection, which links the inputs directly to the outputs of the network layers, is an effective strategy to mitigate the issues of vanishing or exploding gradients that can otherwise impair the network's efficacy. This technique enhances the gradient flow, thus preserving the network's ability to learn efficiently throughout the training process.

3. Research on music genre classification model based on CNN and migration learning.

3.1. Music acoustic feature extraction.

The research steps in music genre classification generally include music acoustic feature extraction, music genre classifier training, and music genre classifier testing. Music is represented in a variety of forms, including sound files stored in MP3 and wav formats as well as structured symbolic representations in the symbolic music storage format MIDI [21]. However, the musical works input to the music genre classifier need to be a uniform representation, so extracting uniform musical acoustic features is the most crucial step in the music genre classification step.

Commonly used for music genre classification of music short-time features, frequency domain features and inverse spectral domain features and other acoustic features. Among them, the short-time features include Zero-Crossing Rate (ZCR), Short-Time Energy (STE); the frequency-domain features include Spectral Center of mass (SC), Spectral Energy (SE), mel frequency, etc., and the main formulas are shown as follows:

$$ZCR = \frac{1}{2}\sum_{m=0}^{N-1}|sgn[x_n(m)] - sgn[x_n(m-1)]| \tag{12}$$

$$STE = \sum_{k=-\infty}^{\infty}[x(k)w(n-k)]^2 \tag{13}$$

$$SC = \sum_{n=1}^{N} f(n) * P(E(n)) \tag{14}$$

$$SE = \frac{1}{N}\sum_{n=1}^{N}|F_t(k)|^2 \tag{15}$$

$$mel(f) = 2595 * log_{10}(1 + f/700) \tag{16}$$

The core of music genre classification lies in the extraction of uniform acoustic features from diverse musical expressions, which include short-time, frequency-domain, and cepstrum-domain features, which together form the basis for training and testing of music genre classifiers.

3.2. Music genre classification model.

In this paper, we propose a modeling framework for music genre classification based on deep convolutional neural network and migration learning. It is shown in Figure 4. The main idea of the model is to first train a DCNN model on the publicly available dataset used in this paper to determine the network weights and give the network feature representation capabilities. This knowledge is then migrated to the fine-grained image recognition task, using these pre-trained networks as feature extractors to obtain a feature representation of the fine-grained image dataset [22]. The process of completing the classification of music genres based on deep convolutional neural networks and migration learning can be divided into four main steps: model pre-training, feature extraction, classifier training, and category probability prediction.

This manuscript introduces a novel framework for categorizing music genres, leveraging a Deep Convolutional Neural Network (DCNN) and transfer learning techniques, as depicted in Figure 4. The essence of our model involves an initial phase of training the DCNN on a publicly accessible dataset to ascertain the optimal network weights, thereby endowing the network with the capacity to represent features effectively [23]. Subsequently, this accrued knowledge is transferred to the realm of fine-grained image recognition, wherein the pre-trained networks serve as feature extractors to derive a representative feature set

from the fine-grained image dataset. The methodology of employing deep convolutional neural networks in conjunction with transfer learning for music genre classification unfolds across four principal stages: the preliminary training of the model, the extraction of discriminative features, the subsequent training of the classifier, and the final computation of genre-specific probabilities.

The weight parameters in the training process of DCNN models are generally randomly initialized first, and a large number of training samples are usually required to make the network adequately trained so as to obtain a better feature representation capability. However, in the music genre classification task, the samples available for training are very limited, and thus cannot meet the training requirements of the DCNN model. Thus, we chose to fully pre-train the model on a large public dataset to determine the network weights, so that the model can obtain the feature representation capability. Second, the completed trained DCNN is used as a generalized feature extractor to obtain the feature representation of the music genre classification dataset, a process called forward conduction.

During the feedforward phase of the DCNN, the starting point is the preprocessed musical dataset designated for genre classification [24]. Subsequently, the network's layered weights are engaged in computations, culminating in a series of feature representations that encapsulate the data's intrinsic characteristics at various layers. It is important to note that the network's weights remain static throughout this operation. For the application of transfer learning, the terminal components of the network— the classification layer along with any subsequent fully connected layers— are typically excised. This modification allows the output from the DCNN's convolutional layers to serve as the extracted feature set, which is then utilized to characterize the attributes of the new dataset effectively.

Next, after obtaining the feature representation of the music genre, a classifier is constructed on the features of the training set, because the features extracted by the DCNN model are already discriminative features, so in order to prevent the model from overfitting, a very complex classifier is generally not used here. Finally, the MLP model, which has been trained on the training set, is predicted on the test data to complete the task of classifying music genres, and the average of the prediction correctness and the average classification accuracy is chosen to evaluate the effectiveness of the model.

4. **Experiments and results.**

4.1. **Data sets and evaluation indicators.**

The common music genre dataset does not include all genre types, but it lays the foundation for the research of music genre classification methods. Open-source music genre datasets mainly include FMAW dataset, GTZAN dataset, and Emotify music dataset. In this paper, we use the FMA dataset as the training set of the music genre classifier, which has the advantages of more data samples and rich genre types compared with other datasets [25].

The FMAW dataset provides 917GB in size and collects 106,574 music samples from 16,341 artists, 14,854 albums, and a total of 161 music genres. The dataset is unevenly distributed across genres, with the rock genre having the highest number of samples at over 7,000, and the lowest number of samples at only 20 for light music [26]. The data samples are stored in MP3 format with a sampling rate of 44,100 Hz, and include data from 161 music genres.

In this paper, the MAP and ACC are used as evaluation metrics for music genre classification models. The mean of average classification accuracy is a commonly used evaluation metric for classification problems to measure how well the model predicts

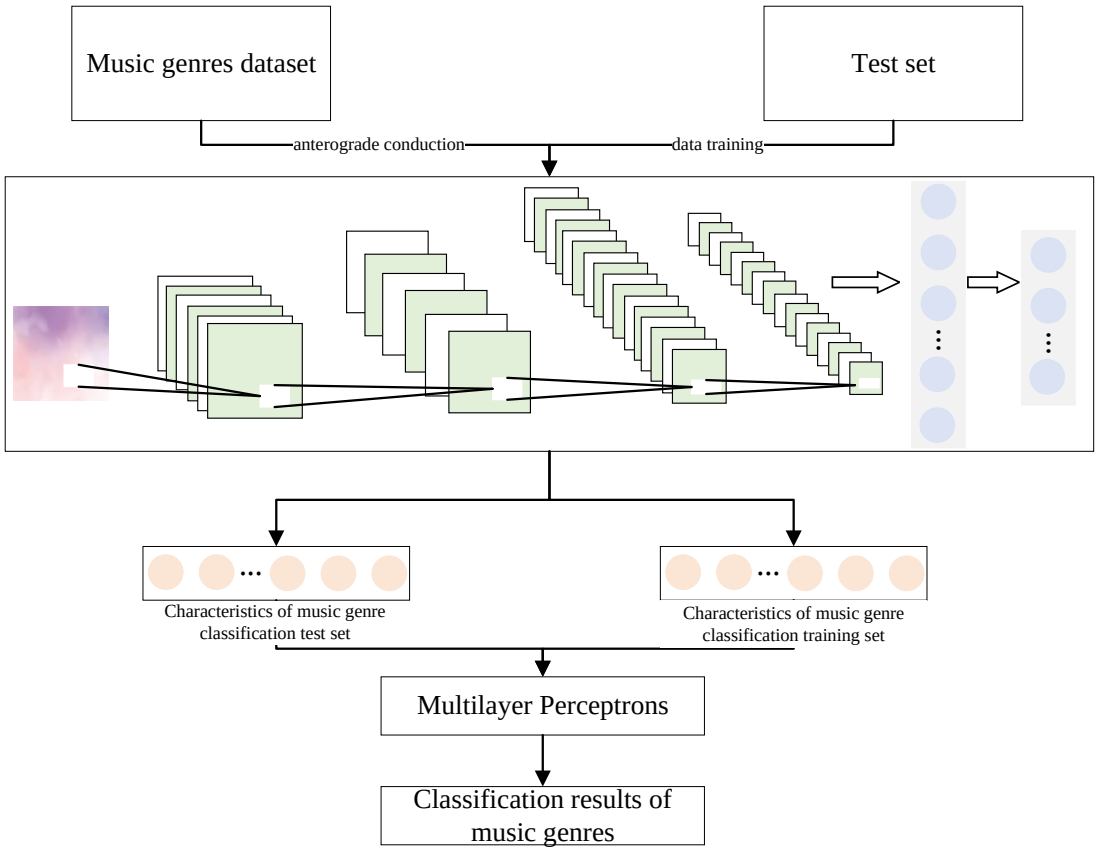

Figure 4. Model framework diagram.

all categories. The prediction accuracy is the percentage of genre categories correctly predicted. The MAP is shown as follows:

$$MAP = \frac{1}{K} \sum_{i=1}^{K} \frac{T_i}{T_i + F_i} \tag{17}$$

where K represents the total count of distinct genre categories present in the test data collection. T corresponds to the aggregate tally of accurate genre classifications made by the model. Conversely, F is utilized to express the total sum of misclassifications, where the model's predictions did not align with the actual genres. The ACC is shown as follows:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{18}$$

where TP denotes instances correctly identified as positive, FP refers to instances mistakenly categorized as positive, TN signifies cases correctly recognized as negative, and FN represents cases erroneously classified as negative. Different quantitative analyses can be performed for different classifications when evaluating methods. For example, in music genre classification identification, TP denotes the number of samples that correctly identify the genre type.

4.2. **Experimental results and discussion.** The objective of this research is to assess the efficacy of our novel music genre classifier, designated CNN-TL, which leverages the power of CNNs coupled with transfer learning techniques. To ensure a comprehensive

benchmark, the CNN-TL's performance is juxtaposed against a suite of prevalent machine learning methodologies recognized within our domain. This comparative analysis encompasses the Decision Tree, Logistic Regression, the ensemble method known as Random Forest, and the SVM, with particular emphasis on its application of the Radial Basis Function (RBF) as the kernel. This kernel selection for the SVM is noteworthy and is reiterated for its significance in the comparative study.

For our experiments, we used the FMAW dataset, which is a widely recognized music genre classification dataset containing audio files from a wide range of music genres. To ensure a fair comparison, all algorithms were trained and tested on the same dataset. For traditional machine learning algorithms, we first manually extracted acoustic features from the audio files, which include musical short-time features, frequency-domain features, and cepstrum-domain features, which are commonly used feature sets in music genre classification. The experimental results are shown in Table 1.

Table 1. Model comparison results.

Classification Model	**MAP**	**ACC**
Decision Tree	0.572	46.92%
Logistic Regression	0.603	59.78%
Random Forest	0.697	61.56%
Radial Basis Function	0.752	67.32%
CNN-TL (Proposed)	**0.843**	**86.34%**

As shown in Table 1, the experimental results show that the CNN-TL model outperforms other comparative algorithms in the average evaluation metrics of classification accuracy and average classification precision. The superior performance of the CNN-TL model is mainly attributed to its deep structure, which is able to efficiently and automatically learn complex music features from the original audio signals. In addition, the application of migration learning further improves the generalization ability of the model, enabling it to adapt to different music genres and maintain a high classification accuracy even in the face of unseen music samples. In contrast, while traditional machine learning algorithms have some flexibility in feature engineering, they are often limited by the expressive power of hand-extracted features when dealing with high-dimensional, nonlinear music data. The CNN-TL model avoids this limitation through an end-to-end learning approach, resulting in better performance in music genre classification tasks. In summary, the CNN-TL model not only performs well in experiments, but also its automated feature extraction and strong generalization capabilities provide advantages for its deployment in real applications. Future work will further explore the potential of the CNN-TL model in larger datasets and more complex music scenes.

5. **Conclusions.** Based on an in-depth analysis of the music genre classification problem, this paper proposes an innovative automatic classification model that combines the advanced feature extraction capability of CNNs with the flexibility of transfer learning techniques. With CNNs, we successfully extract deep features of music signals automatically and enable the pre-trained model to quickly adapt and learn new classification tasks through a migration learning strategy. Experiments on several public datasets validate the high classification accuracy and excellent generalization performance of the proposed model, which achieves a significant leap in both efficiency and accuracy compared with traditional methods. In addition, this study not only opens up a new way of music genre classification, but also provides strong empirical support for the application of deep learning in music information processing. Meanwhile, we plan to further optimize the

model architecture, explore advanced strategies for multimodal data fusion, and commit to applying the model to a wider range of music scenarios, with a view to continuously improving classification accuracy and optimizing users' music experience.

REFERENCES

[1] K. K. Jena, S. K. Bhoi, S. Mohapatra, and S. Bakshi, "A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis," *Neural Computing and Applications*, vol. 35, no. 15, pp. 11223–11248, 2023.

[2] A. Dorochowicz, A. Majdańczuk, P. Hoffmann, and B. Kostek, "Comparison of classification of musical genre obtained by subjective tests and decision algorithms," *Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3725–3725, 2017.

[3] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19–46, 2024.

[4] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, 2019.

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

[6] F. Hu, G. Xia, J. Hu, and L. Zhang, "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14680–14707, 2015.

[7] R. Johnson and T. Zhang, "A Framework of Composite Functional Gradient Methods for Generative Adversarial Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 17–32, 2021.

[8] S. Liao, J. Wang, R. Yu, K. Sato, and Z. Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," *Procedia Computer Science*, vol. 111, no. 1, pp. 376–381, 2017.

[9] A. Jamatia, A. Das, and B. Gambäck, "Deep Learning-Based Language Identification in English-Hindi-Bengali Code-Mixed Social Media Corpora," *International Journal of Intelligent Systems*, vol. 28, no. SP3, pp. 399–408, 2019.

[10] J. C. Pereira, E. R. Caffarena, and C. N. d. Santos, "Boosting Docking-Based Virtual Screening with Deep Learning," *Journal of Chemical Information and Modeling*, vol. 56, no. 12, pp. 2495–2506, 2016.

[11] P. Zhang, W. Li, X. Ma, J. He, J. Huang, and Q. Li, "Feature-selection-based transfer learning for Intracortical brain–machine Interface decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 60–73, 2020.

[12] X. Luo, B. D. Roads, and B. C. Love, "The Costs and Benefits of Goal-Directed Attention in Deep Convolutional Neural Networks," *Computational Brain & Behavior*, vol. 4, no. 2, pp. 213–230, 2021.

[13] A.-A. Liu, H. Du, N. Xu, Q. Zhang, S. Zhang, Y. Tang, and X. Li, "Exploring visual relationship for social media popularity prediction," *Journal of Visual Communication and Image Representation*, vol. 90, 103738, 2023.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[15] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification With Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[18] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[19] R. M. Manshani, K. Suryawanshi, and H. Mathur, "Classification of Music genres using Machine Learning," *International Journal of Scientific Research and Management*, vol. 10, no. 09, pp. 918–927, 2022.

[20] A. Rosner, B. Schuller, and B. Kostek, "Classification of Music Genres Based on Music Separation into Harmonic and Drum Components," *Archives of Acoustics*, vol. 39, no. 4, pp. 629–638, 2014.

[21] L. Wen, L. Gao, and X. Li, "A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2019.

[22] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2020.

[23] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Evolving Deep Convolutional Neural Networks for Image Classification," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 394–407, 2020.

[24] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, 2019.

[25] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7313–7331, 2020.

[26] R. Du, S. Zhu, H. Ni, T. Mao, J. Li, and R. Wei, "Valence-arousal classification of emotion evoked by Chinese ancient-style music using 1D-CNN-BiLSTM model on EEG signals for college students," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15439–15456, 2022.