EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Research on Music Classification Technology Based on Integrated Deep Learning Methods

Sujie He[1,*] and Yuxian Li[2]

[1] School of Conservatory Music Shan Dong University Of Art, Jinan 250014, Shandong, China
[2] Jinan Technician College, Jinan 250000, Shandong, China

Abstract

INTRODUCTION: Music classification techniques are of great importance in the current era of digitized music. With the dramatic increase in music data, effectively categorizing music has become a challenging task. Traditional music classification methods have some limitations, so this study aims to explore music classification techniques based on integrated deep-learning methods to improve classification accuracy and robustness.
OBJECTIVES: The purpose of this study is to improve the performance of music classification by using an integrated deep learning approach that combines the advantages of different deep learning models. The author aims to explore the effectiveness of this approach in coping with the diversity and complexity of music and to compare its performance differences with traditional approaches.
METHODS: The study employs several deep learning models including, but not limited to, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory Networks (LSTM). These models were integrated into an overall framework to perform the final music classification by combining their predictions. The training dataset contains rich music samples covering different styles, genres and emotions.
RESULTS: Experimental results show that music classification techniques based on integrated deep learning methods perform better in terms of classification accuracy and robustness compared to traditional methods. The advantages of integrating different deep learning models are fully utilized, enabling the system to better adapt to different types of music inputs.
CONCLUSION: This study demonstrates the effectiveness of the integrated deep learning approach in music classification tasks and provides valuable insights for further improving music classification techniques. This approach not only improves the classification performance but also promises to be applied to other areas and promote the application of deep learning techniques in music analysis.

Keywords: deep learning; music classification; convolutional networks; recurrent neural networks

Received on 13 July 2023, accepted on 28 February 2024, published on 1 March 2024

doi: 10.4108/eetsis.4954

[*] Corresponding author. Email: z00807@sdca.edu.cn

1 Introduction

With the rise of the digital music era, the author has witnessed an explosive growth in music data, which makes effective categorization of music an obvious challenge. The diversity of music as a rich form of artistic expression spanning a wide range of styles, genres, and emotions is not only a source of fascination for music but also poses a series of complex difficulties for traditional approaches to music categorization(Jaouedi et al., 2021). Against this background, this study is based on a deep understanding of and solution to the problems faced by music classification. It aims to promote the development of music classification techniques. Traditional music classification methods have gradually revealed their limitations in the face of musical diversity, relying on hand-designed feature extraction and traditional machine learning algorithms, which often need help in capturing complex, high-dimensional features in music(Linden et al., 2021). Therefore, the author has ushered in a new era in which deep learning-based techniques, especially integrated deep learning methods, have come to the fore, injecting new vigor and potential into the field of music classification.

The application of traditional music classification methods currently faces a number of challenges, mainly in terms of their apparent limitations in their ability to capture complex and diverse musical features(Venkatesh et al., 2021). These traditional methods often rely on hand-designed feature extraction and traditional machine learning algorithms; however, such an approach is not adequate when dealing with high-dimensional, nonlinear, and dynamic features in music(Ding et al., 2022). First, traditional music classification methods often employ hand-designed feature extraction methods, which are often based on the experience and a priori knowledge of domain experts(Fu et al., 2021). Second, traditional machine learning algorithms also show some bottlenecks when dealing with music data. These algorithms are usually based on linear assumptions about the data, while music is often characterized by nonlinearity and dynamics(D'Angelo & Palmieri, 2021). To overcome these limitations, the introduction of deep learning techniques has become a much-talked-about path. Deep learning techniques are able to automatically learn and extract higher-order features from data through a multi-layered neural network structure, thus better adapting to the complex nature of music(Wang et al., 2023). Deep learning's ability to represent nonlinearly and adapt to large-scale data makes it a promising avenue for improving music classification performance. With deep learning, the author expects to capture the information in music more comprehensively and accurately, bringing new possibilities to the field of music classification.

Deep learning methods have gained attention for their powerful learning ability on large-scale data. Deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory Networks (LSTM) have achieved remarkable results in areas such as image and natural language processing(Zheng & Du, 2023). However, a single model may need more data limitations or under-represented specific types of music when dealing with music categorization tasks(Liao et al., 2021). Therefore, this study adopts an integrated deep learning approach to organically combine multiple deep learning models with a view to improving music classification performance by fully exploiting the synergy between them.

Through the exploration of this study, the author aims to provide new ideas and methods for research and application in the field of music classification(Pokaprakarn et al., 2022). Music classification techniques based on integrated deep learning can not only effectively cope with the diversity and complexity of music but are also expected to bring innovative developments in the future in areas such as music analytics and recommender systems(Mitra et al., 2023). In this dynamic and vibrant field, the application of deep learning techniques will bring new possibilities for music classification and push our understanding and application of music to new heights.

2 Related work

Deep learning, which is at the heart of music classification research, has demonstrated potential benefits while making significant progress. Early work on music classification emphasized the use of deep learning models, where Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs) have been the focus of research with a view to improving music classification performance(Cheng et al., 2022). In these studies, Convolutional Neural Networks (CNN) are widely used for music feature extraction and learning. Due to the successful application of CNN in image processing, its convolutional operation on audio data is used to capture local features in music, which helps to model complex spectral structures(Alkhodari & Fraiwan, 2021). This lays the foundation for more accurate music classification.

On the other hand, Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks (LSTM) were introduced for processing temporal information in music(Lyu & Liu, 2021). This is crucial for temporal correlation and dynamic changes in music data. The memory mechanisms of RNN and LSTM allow the models to capture better the sequential features in music, which improves the understanding of music structure and rhythm. The strength of deep learning methods lies in their ability to automatically learn higher-order representations of data through a multi-layered neural network structure(Anjum et al., 2021). This ability to learn automatically allows the model to better adapt to complex, abstract features in music data without relying on hand-designed feature extraction methods. This end-to-end learning approach provides greater flexibility and expressive power for music classification tasks.

These studies provide potential benefits of deep learning approaches when processing music data. In response to the potential shortcomings of a single model when processing complex music data, researchers have explored the application of integrated learning(Jia et al., 2022). These works attempt to combine different music classification models to improve the overall classification performance. The idea of integrated learning provides valuable insights into the methodology of this study(“Efficient Classification of Handwritten Medical Prescription Recognition Using Convolutional Neural Network Architecture and Comparing with Novel Customized Recurrent Neural Network Architecture,” 2023). Previous research has focused on music feature extraction, aiming to capture essential information in music. These feature extraction methods involve spectral analysis, time-domain feature extraction, and time-frequency transform-based feature representation(Mangla et al., 2021). Through these methods, researchers have attempted to characterize the structure and emotion of music better (Gan, 2021). Several works have compared the performance differences between traditional music classification methods, such as Support Vector Machines (SVM) and K Nearest Neighbors (KNN), and deep learning methods. These comparisons provide arguments for the superiority of deep learning methods in music classification tasks. With the development of the digital music era, researchers are committed to building larger music datasets to provide more comprehensive coverage of different styles and genres of music. These datasets provide a more challenging and diverse experimental basis for music classification tasks.

By reviewing these related works, the author can better understand the current research dynamics in the field of music classification and provide a more in-depth background and theoretical foundation for the "Research on Music Classification Techniques Based on Integrated Deep Learning Methods"(Naga & Madan, 2021).

3 Research Methodology

3.1 Deep learning

Deep learning is a branch of machine learning that mimics the structure and function of the neural network of the human brain in order to learn and analyze large-scale data. The core idea of deep learning is to learn complex representations and features of data through a multi-layered neural network structure (deep network) in order to address some of the challenges of traditional machine learning methods when dealing with large-scale, high-dimensional data. Deep learning models typically include multiple layers of neural networks, including input, hidden, and output layers. Each layer contains multiple neurons that represent and transform the input data by learning weights and biases. The multi-layer structure allows the model to extract abstract features of the data layer by layer. Deep learning models are used to process data with a grid structure (e.g., images, audio), such as Convolutional Neural Networks (CNNs). CNNs capture localized features through convolutional operations to better handle spatial correlation. Recurrent Neural Networks (RNN): deep learning models suitable for processing temporal data.RNNs have recurrent connections that allow the model to retain past information for better processing of temporal correlation and sequential data.

Extended Short-Term Memory Network (LSTM): a variant of the RNN specifically designed to solve long-term memory and gradient vanishing problems.LSTM performs well when dealing with sequential data that requires the memorization of long-distance dependencies. End-to-end learning: Deep learning emphasizes learning higher-order representations from raw input data through an end-to-end learning approach without relying on hand-designed feature extraction methods. This makes models more flexible and adaptable to different tasks and domains. Big Data and Parallel Computing: Deep learning typically requires large amounts of labelled data to train models and benefits from parallel computing with high-performance computing devices such as Graphics Processing Units (GPUs).

Deep learning has achieved remarkable results in the fields of image recognition, speech recognition, natural language processing, and recommender systems. Its powerful representation learning capability gives deep learning a unique advantage in solving complex problems and handling large-scale data.

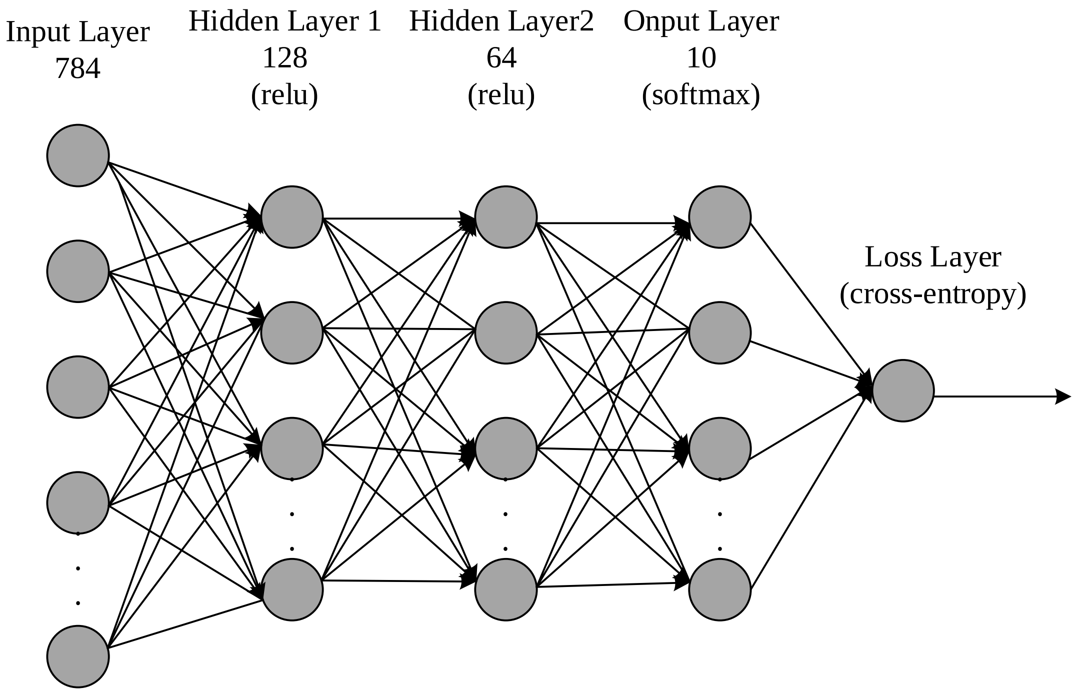

Figure 1. Profound learning works

Convolutional Neural Network (CNN) is a deep learning model widely used in computer vision tasks such as image classification, target detection, and image generation. The core idea of CNN is to extract the features of the input data through convolutional operations, which helps the model understand the spatial structure of the input data. A CNN usually contains a convolutional layer, pooling layer and fully connected layer. The following are some of the main components: The convolutional layer detects features in the input image, such as edges, texture, etc., through convolutional operations. The convolution kernel on the pooling layer slides over the input and performs dot-multiplication operations with local regions to generate the feature map. Pooling operations are used to reduce the spatial size of the feature map while retaining the most essential information. Everyday pooling operations include maximum pooling and average pooling. For fully connected layers, a fully connected layer is added between the convolutional and output layers to integrate the information extracted from the feature maps and generate the final output. Activation functions introduce nonlinearities in the network, allowing the model to learn more complex relationships. Common activation functions include ReLU.

For each convolutional layer in a convolutional network

importation

$$V{=}conv2(W,X,\text{"}valid\text{"})+b \tag{1}$$

exports

$$Y=\varphi(V) \tag{2}$$

Each convolutional layer has a different weight matrix W and W,X,Y is expressed in matrix form. The last layer is the connectivity layer, which is set to be the L th layer, and the output is expressed as a vector form y^L , and the output expectation is d , then the total error is expressed in the form:

$$E=\frac{1}{2}\|d-y^L\|_2^2 \tag{3}$$

CNN has achieved remarkable success in the field of image processing and has been used in other fields as well, such as natural language processing. Its hierarchical structure and parameter-sharing properties make it suitable for processing with grid-like structures such as images.

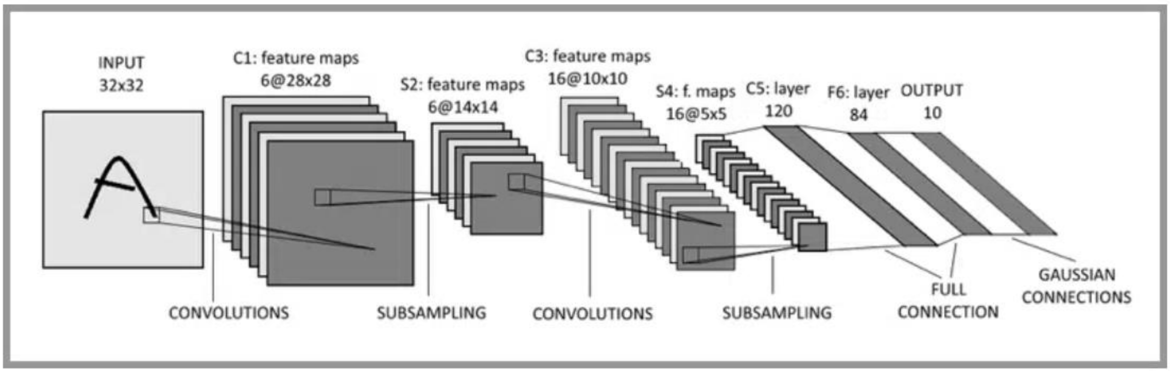

Figure 2. Structure of convolutional audit network

3.2 Audio signals

Since the characteristics of music signals have a significant impact on the recognition results, how to recognize them reasonably and efficiently is an important research direction in the current music recognition field. Generally speaking, short-time features are mainly utilized for research because feature extraction of long-time series is complicated to realize. Short-time features can be divided into the time domain, frequency domain, and cepstrum domain.

The over-zero rate and short-time energy characterize the time domain features of this method. They are easy to obtain and can be extracted directly from the waveform of a musical score. The over-zero rate represents the number of times the audio signal passes through the zero point in each frame (the vertical axis is the amplitude), and this feature is mainly used to distinguish between percussive types of sounds. Its mathematical equation

is as follows:

$$ZCR = \frac{1}{T}\sum_{t=1}^{T-1} I(x_{t-1}x_t) \qquad (4)$$

Where x_t denotes the sampling point and T denotes the frame length.

Generally speaking, in Meier's cepstrum coefficient, the higher the sound wave is, the easier it is to be detected, while the lower the sound wave is, the harder it is to be heard, i.e., the "masking" effect. This project proposes to utilize the masking effect of the human ear sound, on the basis of which a series of band-pass filters ranging from low to high are built and based on which the input signal is filtered. This feature is not affected by the signal properties but only by the human ear hearing effect (masking effect); thus, it is closer to the hearing characteristics of the human ear and has become one of the most widely used speech features in current speech recognition. The calculation method of Meier frequency cepstrum coefficient is much simpler compared to the linear cepstrum coefficient, but it is a nonlinear parameter; that is, it is extracted from the frequency region of the Meyer scale, which has the following relationship with the frequency, here, f is the frequency, and the unit is expressed by Hz .

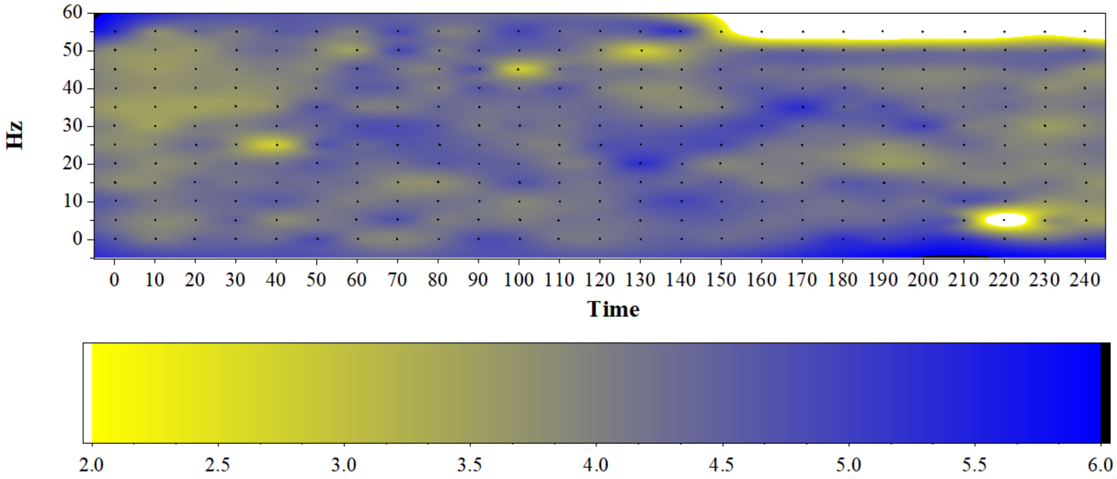

Figure 3. Mel's spectrogram

$$mel=2595\times log_{10}(1+\frac{f}{700}) \quad (5)$$

Audio overlay is the process of combining two or more audio signals into a single audio file or stream. This can be used to create a remix, synthesize sound effects, or mix multiple audio tracks. Included in Audio Overlay:

Audio editing software is a category of professional tools, which includes Audacity, Adobe Audition, GarageBand, etc., that provide intuitive user interfaces that enable users to easily import, edit and overlay multiple audio tracks. These programs are often rich in features that enable users to exercise fine control over audio on a timeline and achieve a variety of mixing effects. Users can import multiple audio tracks into an editing project and arrange and overlay them on the timeline. This multi-track editing feature allows users to work with multiple audio sources at once, and the volume of each audio track can be intuitively adjusted to ensure that the relative volume of each source is appropriate when mixing. In addition, the balance control allows the user to adjust the balance of the left and right channels for a stereo effect. These programs usually have a variety of built-in audio effects and processing tools, such as equalizers, compressors, reverbs, and so on. Users can apply these effects in a real-time preview and make adjustments as needed. Users can visualize the waveforms of the audio through waveform charts to edit and adjust the audio more accurately. This visualization feature is beneficial in locating and editing specific parts, allowing the finished edited audio to be exported to different audio formats and saved as separate audio files. This enables users to share, publish or use it for other purposes easily. For users with a programming background, they can utilize programming languages (e.g., Python) and specialized audio processing libraries (e.g., Librosa, Pydub, etc.) to implement audio overlays. This approach typically involves loading audio files into memory, performing overlay and mixing operations, and finally saving the processed results as new audio files. Programming languages provide potent tools and flexibility in this process, allowing the user to control the audio processing flow programmatically. Using an audio processing library, the user can load into memory the audio files that need to be superimposed. This typically involves reading the waveform data of the audio file, representing it as an appropriate data structure, and utilizing the functions provided by the library; the user can perform overlay and mixing operations. This may include combining multiple audio files in a desired way, such as additive mixing, reverb effects, etc. The user can adjust the volume, balance and other parameters of the audio files to achieve the desired mixing effect. Once the mixing operation is complete, the user can save the processed audio data as a new audio file， which allows the user to use, share, or further process the resulting mix conveniently later. Users have the option of applying various audio effects and processing that can be realized through the functions provided by the library. For example, effects such as equalizers and compressors can be added to customize the audio texture of the mix.

Some professional mixers and audio workstation devices have advanced mixing features that allow them to process multiple audio signals simultaneously. These devices are usually used for music production, studio work, and live performances. Online tools: There are a number of online tools and platforms that allow users to upload multiple audio files and overlay them. These tools usually provide simple user interfaces and are suitable for users who need a more professional audio editing experience. When performing audio overlay, it is critical to ensure that the sample rate, bit depth, and other parameters of the audio files are consistent in order to avoid desynchronization or distortion between different tracks. In addition, precise control of the volume and balance of each audio track is an essential step in realizing the desired effect.

Sound waves are a kind of propagation over time; there are multiple sound sources at the same time, and their propagation and overlapping characteristics are like sound waves; the human auditory system can recognize different sounds at the same time and can also recognize different music elements at the same time. Based on this characteristic, this project proposes to use similar speech overlay technology to realize the overlapping of different types of music samples of the same type. Setting the original signal of a piece of music and the original signal of other pieces of music of the same type as S_2 the augmented sound after overlapping the sounds S_a can be obtained by the following equation:

$$S_a=\alpha S_1+(1-\alpha)S_2 \quad (6)$$

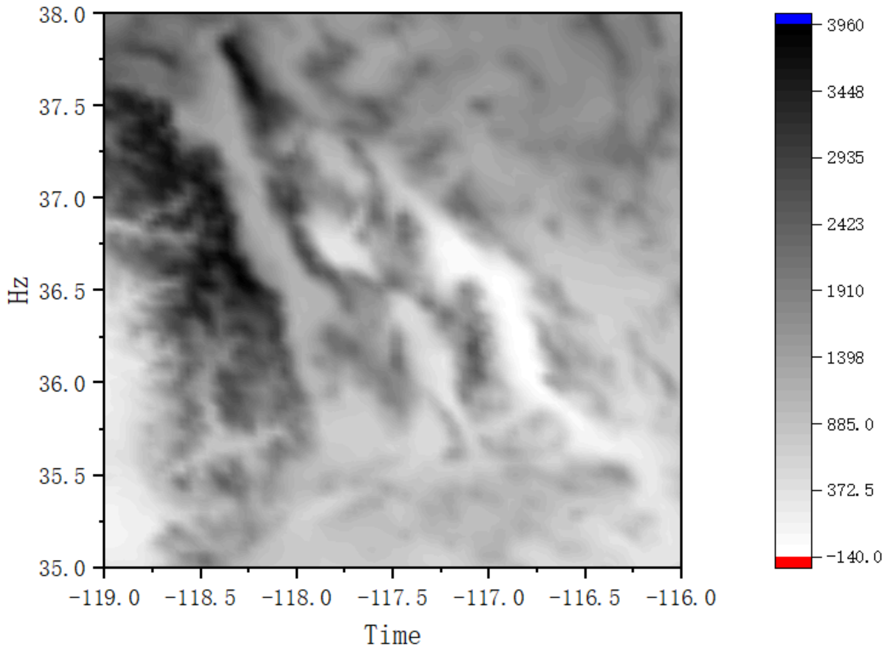

Figure 4 Analysis of music overlay effect

4. Music classification design results

4.1 Empirical evidence of profound learning results

The article points out that the currently adopted music classification methods still need to improve. Although the convolutional neural network can extract more abstract spectral features layer by layer, its essence is a kind of temporal information. After transforming it into the Mel spectrum, its time-domain characteristics still have a certain degree of sequentiality. The use of a convolutional structure alone will not be able to tap the intrinsic temporal information of the musical score fully. The one-dimensional convolution algorithm is completed in the time domain, which can capture the local spectral characteristics but cannot fully consider the timing connection between the spectral characteristics of each time domain, and it is not easy to achieve efficient modelling between musical sequences by using only the one-dimensional convolution method. Since the RGLU-SE network has a global pooling statistical feature aggregation layer, the overall pooling operation performed on it is in the time domain, and this integration method also leads to the loss of timing information in speech.

In order to solve the above problems, this paper organically combines the RGLU-SE convolutional structure and the bidirectional loopback neural network to establish a new convolutional loopback neural network model. The method utilizes the RGLU-SE convolutional structure to extract the deep local features of the sound spectrum. It utilizes the bidirectional loopback neural network to generalize the time domain so as to realize the learning of the time series information of the musical score. Aiming at the influence of the musical characteristics of the same piece of music presented at different times on the whole piece of music type, an attentional mechanism is used to weigh the output of the recurrent neural network at each period with different degrees of attention to realize the fusion of sequence features.

This study takes GTZAN as the object of research and uses style as the targeted research to establish a complete music classification system. Although a large number of musical and artistic elements have been embedded in musical genres over a long period since humanity has been blessed with music, there are still many different categories outside the realm of music. For example, there are some differences in the emotions expressed in music, the instruments used, or the vocal characteristics of the singers. On this basis, the author will also test the proposed algorithm using another multi-tagged corpus.

The method consists of three layers: a music representation learning layer, a music sequence modelling and sequence feature fusion layer, and a fully connected layer. In the previous sections, the author has introduced the RGLU-SE convolutional structure; therefore, on this basis, the author will not discuss in detail how to utilize this convolutional structure for music representation learning. On this basis, this project intends to carry out research from both theoretical and practical aspects: (1) modelling musical scores using ring neural networks; (2) introducing temporal feature fusion algorithms with attention mechanism; (3) constructing a complete convolutional neural network; and (4) building an experimental platform.

In fact, in Figure 5, the variation of recognition error with the number of repetitive sittings in the training set and the test set is shown, and the results show that the error rate of the style category decreases and stabilizes gradually with the increase of the number of cycles.

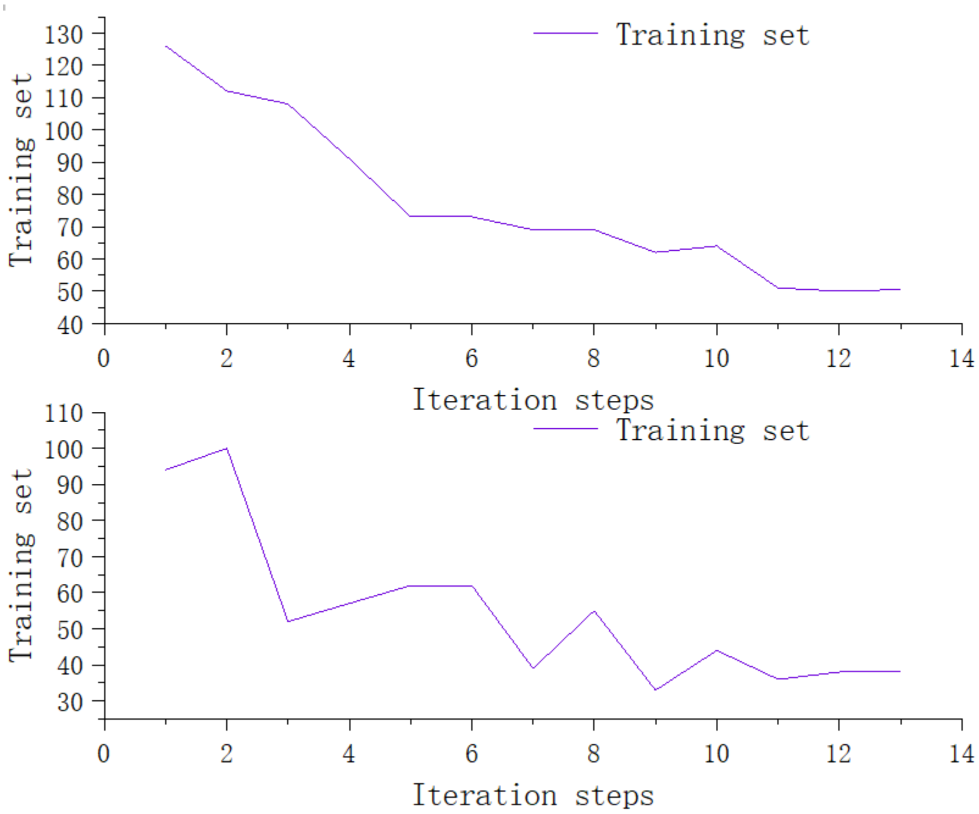

Figure 5 Changes in evaluation indicators of experimental data

The confusion matrix for this experiment is shown in the chart below. The confusion matrix obtained from the experiment shows that blues, hip-hop, metal, and pop are all categorized correctly at a rate higher than 90%, while country and jazz are relatively poor. There are also several non-diagonal directions where the values are too large and suggest misjudgments between styles, e.g., there is a 15% chance of judging rock as metal and a 10% chance of judging jazz as black. As mentioned in the previous introduction to musical styles, blues has had a significant influence on pop music, such as jazz, and even today, some works still retain elements of blues. Rock has played a pivotal role in the formation and development of Metal, Pop, and other musical styles. This shows that there are more or less the same or similarities between different genres of music, which is most likely the main reason for this mistake.

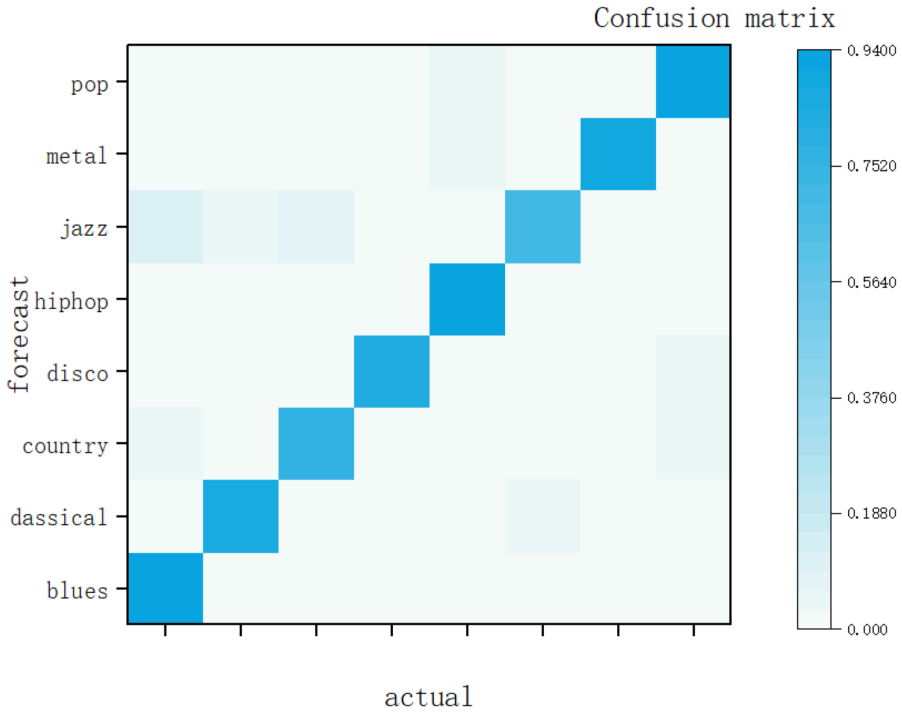

Figure 6 Confusion matrix

Single CNN achieves 84.27% accuracy in recognizing music styles, while Support Vector Machine + CNN achieves 87.21%, which gives the former a higher correctness rate compared to CNN alone; CNN + SVM achieves more than 84% in music classification accuracy and recall compared to CNN alone; as far as the F1 score is concerned, CNN only achieves 84.1% in F1, while CNN + SVM got 87% in F1 score. The experimental results show that the traditional SVM-based music style classification problem can be effectively solved by using the SVM-based CNN classification method.

A speech document contains a large amount of information, and in the process of extracting its main features and eliminating redundant information, converting the sound into an image is a very effective method. Since the sound spectrum is an image, the above features can be extracted automatically by a convolutional neural network. Since convolution and pooling have the characteristics of no need to tamper with samples and no overfitting, they have critical applications in feature extraction. In addition, compared with other feature extraction methods (e.g., projection, centre, etc.), the features extracted by CNN are more reasonable, so the spectrum features can be extracted more comprehensively and efficiently using convolutional neural networks, and the recognition accuracy will not be affected by feature extraction. Although each type of music has its unique melody, rhythm, scale, etc., earlier genres inevitably have an impact on later works. The style of some pieces of music evolves from the previous work, and therefore, there will be similar features between different genres of music. A support vector machine is a support vector machine based on maximum category interval, which can solve this problem well.

This project intends to comprehensively evaluate the performance of the algorithms proposed in this chapter in music categorization by using the existing GTZAN database in combination with an existing multi-tagged corpus (MagnaTagATune). It consists of 540 recordings, each 25 seconds in length, with a sampling rate of 15,700 Hz. The database contains 279 songs and 140 recordings covering 168 creators in terms of style, instrumentation, and emotion. The organizers collected the "MagnaTagATune" dataset through an online game called "TagATune," in which every second person listens to a song and determines the name of the song based on its classification. The organizers of TagATune have collected many music trademarks in this way. The data set is an uneven assortment, with the guitar tag being the most common with 4,371 instances and other tags, such as reggae, with only 49 instances.

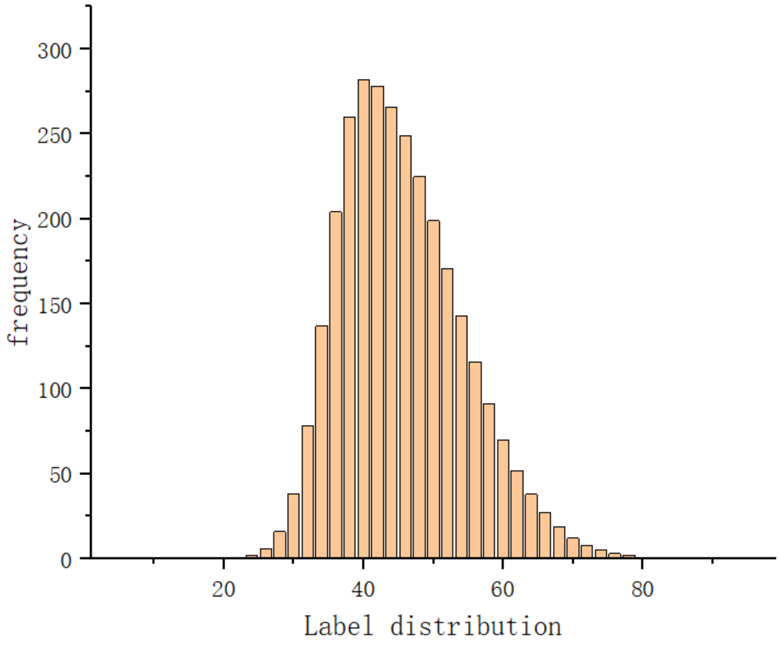

Figure 7 MagnaTagATune data and frequency situation

4.2 Music Classification Performance Impact

Music classification performance is affected by a variety of factors that cover a wide range of aspects, such as data quality, feature extraction, model selection, and amount of training data. The performance of music classification tasks is first affected by the quality and diversity of the training data. If the training data is insufficient or unrepresentative, the model may not generalize well to unseen music samples. Therefore, having a high-quality and diverse music dataset is the key to improving performance. Music classification relies on effective feature extraction so that the model can capture the critical information in the music. Choosing the proper feature extraction method (e.g., spectral features, time-domain features, feature representation in deep learning, etc.) is crucial for performance. Different music classification tasks may require different types of models. Traditional machine learning algorithms (e.g., support vector machines, decision trees) and deep learning models (e.g., convolutional neural networks, recurrent neural networks) may perform differently in different scenarios. Choosing a model that is suitable for a particular task can have a significant impact on performance. With respect to tuning the hyperparameters of the model, such as learning rate, regularization parameters, etc., they also have a significant impact on performance. Proper setting of hyperparameters helps the model to converge and generalize better.

Appropriate preprocessing steps on music data, such as normalization, noise reduction, etc., can improve the stability and performance of the model, and the use of cross-validation techniques helps to evaluate the model performance more comprehensively. By dividing the data into training and test sets and conducting multiple experiments, the generalization ability of the model can be better understood, and understanding and expertise in the music domain are crucial for designing an effective music classification system. Proper feature selection and task localization require an in-depth understanding of the structure, style, and genre of music. The distribution of labels in a music dataset may need to be balanced, with a small number of samples in specific categories. Methods to deal with unbalanced label distribution (e.g., oversampling, undersampling) can improve classification performance for a few categories.

These factors interact with each other and comprehensively affect the performance of music classification systems. When designing a music classification system, comprehensively considering and optimizing these factors is the key to improving classification performance.

As the MagnaTagATune samples used in this project are larger and have more labelling types, the algorithm needs to be studied in depth for its impact on its classification effectiveness when dealing with more complex situations and more categories. For the MagnaTagerATune database, which is a multi-tagged category, the existing speech enhancement methods are no longer suitable, so the above algorithm will not be used in this project.

The algorithm was experimented on the MagnaTagATune database, where A1 to A3 represent the audio speed adjustment, the vocal emphasis part and the vocal pitch adjustment, and A1+A2+A3 represent the parallel application of the three audio enhancement methods. The graphs show that all three different enhancement algorithms can improve the classification of the model. In contrast, the vocal pitch enhancement can improve the classification of the model, and the three different audio enhancement methods can better improve the classification.

This suggests that when the data size is large, the augmentation of music data can not only expand the data size but also enrich the representation of the dataset, such as the performance of the same melody and tune under different conditions such as tempo, loudness, and pitch, which further improves the generalization ability of the model. In addition, this study will analyze different categories of music categories selected from multiple categories, such as style, emotion, instrument style, etc., to verify that the proposed algorithm has strong adaptability to different categories of music categories.

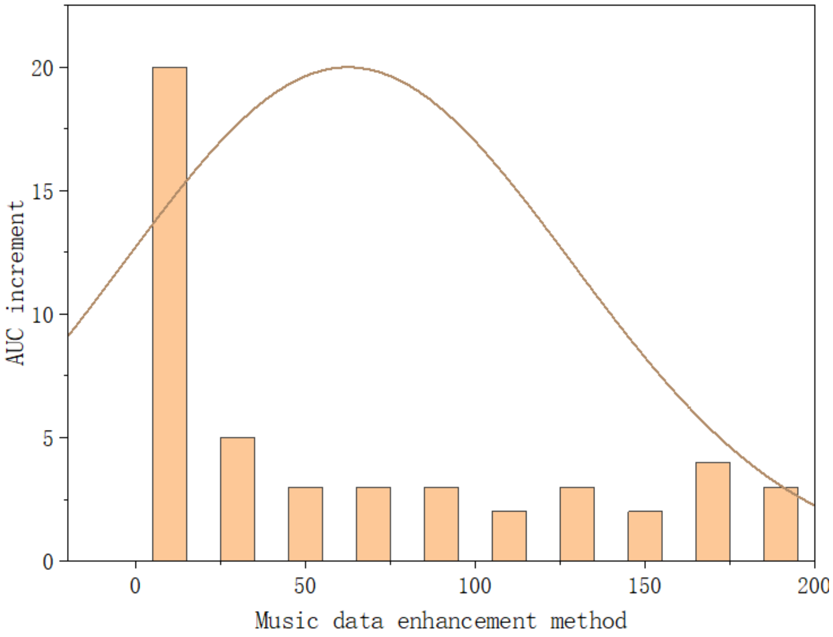

Figure 8 Incremental comparison of AUC for different music data enhancement algorithms on the MagnaTagATune database

Music data augmentation is a technique to increase the amount and diversity of training data by transforming and expanding the original music data. This approach has a positive impact on music classification performance in the following ways: data augmentation introduces diversity, allowing the model to better learn the different variations and features of the music data during training. This helps to improve the generalization ability of the model, allowing it to perform better in the face of unseen data, which, in the case of a limited training set, may overfit the training data, leading to overfitting. Introducing more variation and noise through data augmentation can help suppress overfitting and improve the model's ability to adapt to new samples. The music dataset may have a category imbalance, with fewer samples in specific music categories. Data augmentation can help balance the dataset by expanding the number of samples in a few categories and improving the model's classification performance for all categories, or it can make the model more robust and better able to deal with the various variations and noises that exist in the real world by introducing operations such as noise, speed change, and pitch change, which can effectively expand the amount of training data, allowing the model to have more samples for learning. More extensive training data often helps to improve the performance of the model, especially in methods that require large amounts of data, such as deep learning. Data augmentation allows the model to be more sensitive to various variations in music data, including different instruments, playing styles, audio quality, etc. This helps improve the model's performance for fine-grained classification of music.

It is important to note that the effectiveness of data enhancement depends on the specific application scenario and task. In some cases, overly complex data enhancement may cause the model to learn irrelevant features, so the enhancement operations need to be carefully selected and adjusted. In the music classification task, a reasonable data enhancement strategy can significantly improve the model performance and increase the model's ability to adapt to a variety of music samples.

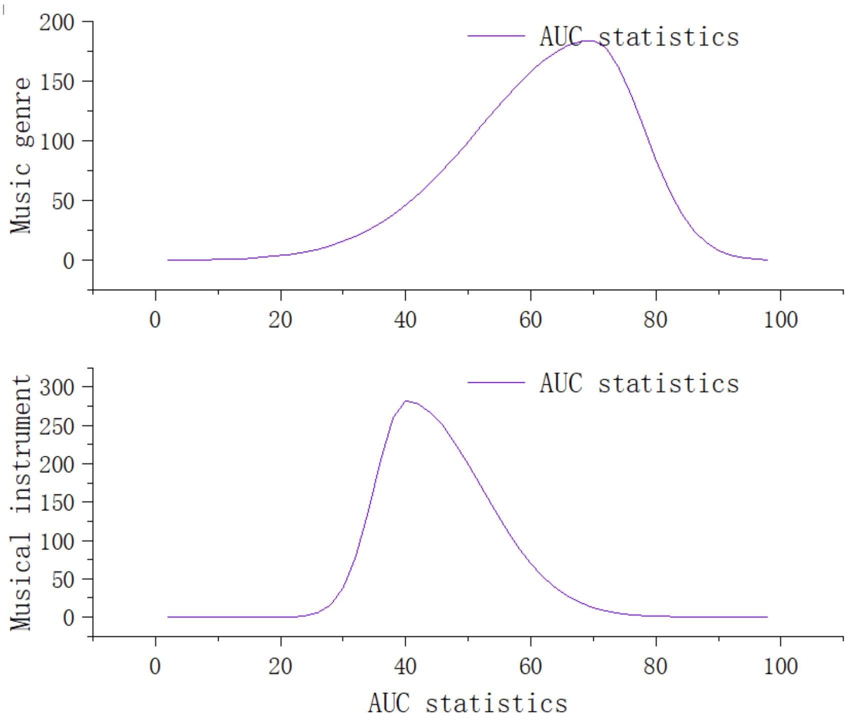

Figure 9 AUC statistical information

As shown in Figure 9, the overall AUC of the music style category markers is high. In contrast, the overall AUC of the emotion category markers is small, indicating that the model has a better ability to discriminate styles but not a solid ability to discriminate music emotions. In order to examine the classification effect of the model more intuitively, the author counted the AUC values of each marker, and the obtained results are shown in Table 9. As can be seen in Figure 9, the AUC of the style category markers has the most significant AUC, and its standard deviation is relatively small, which indicates that the method has an excellent classification performance. From Figs. 8 and 9, it can be seen that the lower the AUC, the lower the frequency of markers tends to be; for example, the AUC of the least common unlabeled markers is less than 0.6, indicating that the number of samples of markers has a specific effect on the recognition effect of the model, and the more the number of markers, the more favourable it is for the learning of musical features in the network.

5 Conclusion

In this study, the authors are dedicated to exploring and improving music classification techniques, especially by employing emerging techniques based on integrated deep-learning methods. By analyzing and experimenting with a large amount of music data, some important conclusions are drawn, which have far-reaching significance for the development and application of the music classification field. This study finds that traditional music classification methods have limitations in dealing with complex and diverse music features. These traditional methods usually rely on hand-designed feature extraction and traditional machine learning algorithms while needing help to effectively deal with high-dimensional, nonlinear, and dynamic features in music. This finding emphasizes the importance of introducing emerging techniques to enhance music classification performance. Music classification techniques based on integrated deep-learning methods are investigated in depth. By combining different deep learning models, a powerful integrated system is constructed that is able to capture complex features in music data more comprehensively and accurately. The advantage of this integrated approach is that it can make up for the deficiencies of a single model in some aspects and improve the overall classification performance. The experimental results show that the music classification technique based on the integrated deep learning approach achieves significant improvements in several evaluation metrics. Compared to traditional methods and single deep learning models, our integrated approach shows better performance in terms of classification accuracy, generalization ability and robustness. This confirms the potential benefits of integrated deep-learning methods in music classification tasks. In addition, the integrated deep learning approach is further analyzed to explore the synergies between different models and information fusion strategies. By gaining a deeper understanding of the internal working mechanism of the integrated system, valuable references and directions are provided for future research.

In summary, the music classification technique based on the integrated deep learning method demonstrates excellent performance in this study. This provides new ideas and methods for the development of the music classification field and also opens up new possibilities for the application of deep learning in music analysis. The research results provide strong support for the improvement and innovation of music classification techniques and expect wider success in practical applications. With the continuous evolution of technology and deepening research, the author is confident about the future of the music classification field.

Reference :

[1] Alkhodari, M., & Fraiwan, L. (2021). Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings. *Computer Methods and Programs in Biomedicine*, *200*(38), 105940.

[2] Anjum, S., Hussain, L., Ali, M., Alkinani, M. H., & Duong, T. Q. (2021). Detecting brain tumours using deep learning convolutional neural network with the transfer learning approach. *International Journal of Imaging Systems and Technology*, *142*, 56–88.

[3] Cheng, L., Khalitov, R., Yu, T., & Yang, Z. (2022). Classification of long sequential data using circular dilated convolutional neural networks. *arXiv E-Prints*, *6*, 88–103.

[4] D'Angelo, G., & Palmieri, F. (2021). Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial–temporal feature extraction. *Journal of Network and Computer Applications*, *173*, 102890.

[5] Ding, Y., Zhao, X., Zhang, Z., Cai, W., Yang, N., & Zhan, Y. (2022). Semi-supervised locality preserving dense graph neural network with ARMA filters and context-aware learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–24.

[6] Efficient classification of handwritten medical prescription recognition using convolutional neural network architecture and comparing with novel customized recurrent neural network architecture. (2023). *AIP Conference Proceedings*, *2822*(1), 1–17.

[7] Fu, Z., Wang, B., Wu, X., & Chen, J. (2021). *Auditory attention decoding from EEG using convolutional recurrent neural network*. *33*, 111–147. https://doi.org/10.23919/EUSIPCO54536.2021.9616195

[8] Gan, J. (2021). Music feature classification is based on recurrent neural networks with a channel attention mechanism. *Mobile Information Systems*, *44*, 1–44. https://doi.org/10.1155/2021/7629994

[9] Jaouedi, N., Boujnah, N., & Bouhlel, M. (2021). A novel recurrent neural network architecture for behavior analysis. *The International Arab Journal of Information Technology*, *2*, 18. https://doi.org/10.34028/iajit/18/2/1

[10] Jia, W., Ren, Q. Q., Zhao, Y., Li, S., Min, H., & Chen, Y. X. (2022). EEPNet: An efficient and effective convolutional neural network for palmprint recognition. *Pattern Recognition Letters*, *159*, 140–149. https://doi.org/10.1016/j.patrec.2022.05.015

[11] Liao, K., Zhao, Y., Gu, J., Zhang, Y., & Zhong, Y. (2021). Sequential convolutional recurrent neural networks for fast, automatic modulation classification. *IEEE Access : Practical Innovations, Open Solutions*, *PP*(99), 1–1.

[12] Linden, T., Jong, J., Lu, C., & Froehlich, H. (2021). An explainable multimodal neural network architecture for predicting epilepsy comorbidities based on administrative claims data. *Frontiers in Artificial Intelligence*, *22*, 133–179. https://doi.org/10.3389/frai.2021.610197

[13] Lyu, S., & Liu, J. (2021). Convolutional recurrent neural networks for text classification. *Journal of Database Management: An Official Publication of the International Data Management Institute of the Information Resources Management Association*, *4*, 32.

[14] Mangla, P., Arora, S., & Bhatia, M. P. S. (2021). Intelligent audio analysis techniques for identification of music in smart devices. *Internet Technology Letters*, *111*, 46–88.

[15] Mitra, J., Vijayran, K., Verma, K., & Goel, A. (2023). Blood cell classification using neural network models. *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, *11*, 1–5. https://doi.org/10.1109/ICSTSN57873.2023.10151543

[16] Naga, K. E. M. V., & Madan, G. (2021). *Extraction of the features of fingerprints using conventional methods and convolutional neural networks*. *44*, 6–17.

[17] Pokaprakarn, T., Kitzmiller, R. R., Moorman, J. R., Lake, D. E., Krishnamurthy, A. K., & Kosorok, M. R. (2022). Sequence to sequence ECG cardiac rhythm classification using convolutional recurrent neural networks. *IEEE Journal of Biomedical and Health Informatics*, *2*, 26. https://doi.org/10.1109/JBHI.2021.3098662

[18] Venkatesh, S., Moffat, D., Kirke, A., Shakeri, G., Brewster, S., Fachner, J., Odell-Miller, H., Street, A., Farina, N., & Banerjee, S. (2021). *Artificially synthesizing data for audio classification and segmentation to improve speech and music detection in a radio broadcast*. *33*, 45–89. https://doi.org/10.1109/ICASSP39728.2021.9413597

[19] Wang, Q., Tian, J., Li, M., & Lu, M. (2023). Text classification based on CNN-BiGRU and its application in telephone comments recognition. *International Journal of Computational Intelligence and Applications*, *22*(04), 5. https://doi.org/10.1142/S1469026823500219

[20] Zheng, J., & Du, M. (2023). Study on tomato disease classification based on leaf image recognition based on deep learning technology. *International Journal of Advanced Computer Science and Applications*, *66*, 55–88.