Hindawi
Scientific Programming
Volume 2022, Article ID 4068207, 6 pages
https://doi.org/10.1155/2022/4068207

Research Article

Application of Music Industry Based on the Deep Neural Network

Minglei Fan

Zhoukou Normal University, Music & Dance Deparment, Zhoukou, Henan 466000, China

Correspondence should be addressed to Minglei Fan; 20081038@zknu.edu.cn

Received 12 October 2021; Revised 9 December 2021; Accepted 13 December 2021; Published 19 January 2022

Academic Editor: Rahman Ali

After entering the digital era, digital music technology has prompted the rise of Internet companies. In the process, it seems that Internet music has made some breakthroughs in business models; yet essentially, it has not changed the way music content reaches users. In the past, different traditional and shallow machine learning techniques are used to extract features from musical signals and classify them. Such techniques were cost-effective and time-consuming. In this study, we use a novel deep convolutional neural network (CNN) to extract multiple features from music signals and classify them. First, the harmonic/percussive sound separation (HPSS) algorithm is used to separate the original music signal spectrogram into temporal and frequency components, and the original spectrogram is used as the input of the CNN. Finally, the network structure of the CNN is designed, and the effect of different parameters on the recognition rate is investigated. It will fundamentally change the way music content reaches music users and is a disruptive technology application for the industry. Experimental results show that the proposed recognition rate of the GTZAN dataset is about 73% with no data expansion.

1. Introduction

Digital music refers to music that has taken shape along with the development of the Internet and is stored and streamed in digital format. In the past, music was mostly stored and distributed in solid form, from the earliest vinyl records to the increasingly lightweight and portable formats of cassette tapes and CD records. Emerging companies with new technologies can often disrupt the industry for some time and become the dominant force in the industry. In the sheet music era, advances in printing technology made lyricists the core of the industry. In the recording era, technology such as phonographs, tapes, and CDs created a large number of technology-based companies. In December 2019, Beijing promulgated the “Implementation Opinions on Promoting the Prosperous Development of Beijing’s Music Industry,” which corresponded to the formation and development of the digital music industry. The capital city of Beijing gathers the best resources of music enterprises, talents, technology, scientific research, and education in the country. The promulgation and implementation of this opinion are of great strategic significance for China’s digital music industry. Facing the major trend of technological integration with the rapid development of new technologies such as 5G, artificial intelligence, and blockchain, it is highly relevant to fully understand the application of blockchain technology in the music industry.

As the music industry environment changes and the threshold of music production decreases, several musicians want to independently produce and distribute their songs; however, in the whole industry process, record companies, publishers’ houses, and major digital music platforms divide a large amount of profits by channels because they occupy a large number of channels and distribution resources. A young music startup team, despite having higher quality music, has no bargaining power in the face of distribution platforms.

The founder of Ujo music has issued a statement saying “In the digital world, is not there a way for music creators to receive royalties directly, instead of going through a slow, inefficient and opaque chain of copyright management?” Is not there a way in the digital world for music creators to receive royalties directly, rather than through a slow, inefficient, and opaque chain of copyright management?” Artists need a platform for free trade and freedom from the domination of a few giant corporations. The application of blockchain technology may provide a prescription for saving artists. Blockchain technology' blockchain distributed recording method is decentralized and cannot be tampered with by anyone, so music creators do not need to go through intermediaries to be able to register their copyrights, and artists can directly distribute their music and receive revenue.

Music style classification is a very challenging but promising task in the field of music information retrieval (MIR) [1]. MIR is the interdisciplinary discipline of retrieving information from music. MIR is a trivial but rising area of research with several real-world applications including signal processing (SP), academic music study (AMS), informatics, machine learning (ML), musicology, psychoacoustics, psychology, computational intelligence (CI), optical music recognition (OMR), and some mixture of these. Since music is an evolving art and there are no clear boundaries between musical styles, automatically classifying musical styles is a challenging problem. The key to the music style classification problem is the feature extraction of musical information. Various feature extraction and classification methods have been proposed in recent years [2, 3], and the performance of these classifiers is highly dependent on the appropriateness of the empirically selected manually extracted features.

Generally, feature extraction and classification in recognition tasks are two separate processing stages, but this study integrates these two stages to better achieve the interaction between the information. Recently, deep CNNs [4] have been making significant progress in general purpose visual recognition tasks [5, 6]. This has stimulated interest in CNNs for classification models [7, 8]. CNNs include multilevel processing of input images, extraction of multiple layers, and high-level feature representation. By sharing some basic components, many manually extracted features and corresponding classification methods can be considered as an approximate or special CNN; however, these features and methods must be carefully designed and integrated to retain discriminative information. Motivated by the remarkable success of CNNs for general purpose visual recognition tasks, this study uses CNNs for the challenging task of music style recognition and investigates the effect of tuning the network structure parameters on recognition rates.

2. Related Work

In this section, we have tried to showcase different classes of recognition and classification by using deep learning (DL) models in the field of computer vision including object recognition, human activity recognition, face recognition, action recognition, text recognition, and many others [9–17]. These techniques have much more dissimilarities such as feature engineering, model selection, extraction method, parameters setting, and cost compared to the music classification and recognition model; however, the music recognition models have specific inspiration from above the models. In this direction, Lee [18] was the first to apply DL to music content analysis, specifically style and artist recognition, by training a convolutional deep confidence network with 2 hidden layers, the convolutional deep belief network (CDBN) in an unsupervised manner to try to activate the hidden layers and generate meaningful features from the preprocessed spectrum. Compared with those standard Mel frequency cepstral coefficients (MFCCs) features, its DL features have higher accuracy. For music style recognition, the music data are transformed into MFCC feature vectors and fed into a CNN with three hidden layers that automatically extract image features for classification, which shows that the CNN has a strong ability to capture the changing image information features. In this study, the following aspects of work have been carried out: (1) using the HPSS algorithm to extract the harmonic and impact components of the music signal spectra in terms of time and frequency, respectively, and use them as the input of the CNN together with the original spectra; (2) designing a CNN-based deep classification framework for music style classification in detail and studying and experimentally demonstrating several key factors required to effectively train a reliable CNN; (3) the CNN-based deep classification framework for music style classification is designed in detail, and (4) this study expands the dataset using affine projection of spectral images and principle component analysis (PCA) to change the pixel values of RGB channels in the training images.

In [19–23], using characteristics of sound as features and experience measures as labels for those features, numerous groups have tried to form ML procedures that can predict emotional reactions based on the sound characteristics alone, usually according to a valence-arousal circumflex manner. Haruvi et al. [24] studied the effects of sound on human focus levels by noninvasive brain decoding expertise. To obtain enhanced thoughtful of the optimal acoustical environment for growing attention levels in listeners, they joint custom app, convenient brain measuring headbands, and ML algorithms to effectively find high temporal resolution focus dynamics from applicants at home. Use the brain decoded focus dynamics; they then examined how numerous properties of sound affected focus levels in diverse jobs.

3. Proposed Work

In this section, the proposed work is discussed.

3.1. Music Style Recognition Algorithm. The general framework of music style recognition based on the CNN is shown in Figure 1. First, the HPSS algorithm is used to separate the music tracks, and the original tracks are separated into harmonic and percussive sources; then, the short-time Fourier transform is applied to these two sources and the original tracks, and the transformed spectra are input into the CNN for learning, training, and prediction, and the final output result is the final recognition rate.

3.2. Harmonic/Percussive Separation Algorithm. Music signals are usually composed of harmonic and percussive sound components, which have very different

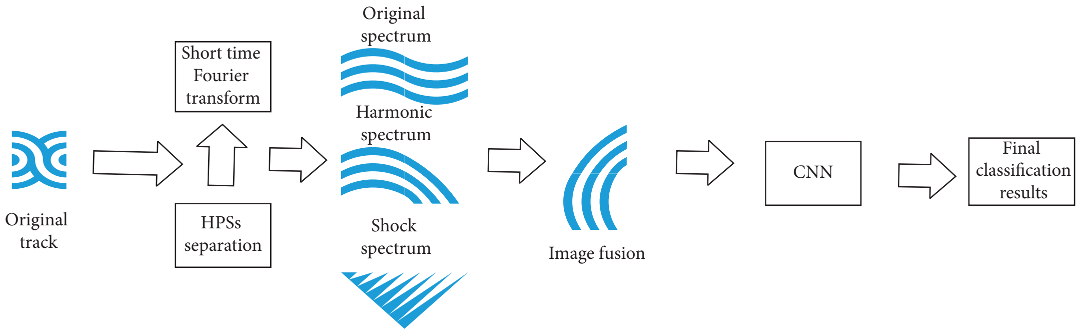

Figure 1: Flowchart of music style recognition.

characteristics. This study uses the harmonics/percussion separation algorithm to separate the harmonic and percussive components of a music signal, which is essentially a signal separation based on the anisotropic continuity of the spectrogram. The key to this method is its focus on the difference between the harmonic and the percussion spectra in the continuum direction. The harmonic spectrum is usually continuous in the time direction, and the shock spectrum is continuous in the frequency direction. Figure 2 shows the original spectrum of a music track and the separated harmonic spectrum and shock spectrum. From the figure, it can be seen that the separated harmonic spectrum is smoothly distributed continuously along the time axis at a fixed frequency, while the impact spectrum is a harmonic sound that strikes in the time axis with the lateral.

3.3. Structure. The network structure of this study is a layered CNN that extracts local features through convolution input images and a group of kernel filters. The convolution layer generates a feature map through a linear convolution filter and nonlinear activation function (ReLU). The output of neurons in the same layer forms a plane, which is called a feature map; then, the convolution feature map is obtained by pooling and filtered to the next layer. Different feature maps are obtained by setting different kernel filters in the local receptive field. Given, X_l^q represents the p^{th} feature map in the first layer, the convolution of the whole feature map and the activation function is represented by

$$X_l^q = \max\left(0, \sum_{X^p \in M_q} X_{l-1}^p \otimes k_l^{pq} + b_l^q\right), \tag{1}$$

where X_l^q is the feature map output from the q^{th} convolution kernel at layer l, ⊗ represents the convolution operation, k_l^{pq} is the convolution kernel, M_q denotes the set of feature maps, X_{l-1}, max(-) is the nonlinear activation function ReLU, b_l^q is the bias, and the feature map k_l^{pq} uses the activation function after the convolution operation.

Since it is found that local response normalization helps to generalize the network, normalization is performed after ReLU in some layers of this network model. This response normalization implements a form of lateral inhibition found in real neurons, and the effect of this lateral inhibition is to make the output values of neurons computed by different convolution kernels more sensitive to the activity of neurons with larger computational values between. The pooling layer uses max pooling, in which both the convolutional and pooling layers alternate in a CNN. The output layer is fully connected to the previous layer, and the feature vectors it generates can be sent to the logistic regression layer to complete the recognition task, and the weights in all networks are learned using a backpropagation algorithm [25].

In this study, we use the stochastic gradient descent (SGD) algorithm (Figure 3) to train the network model and find that small weight decay is very important for the learning of the model. The weight decay can reduce the training error of the model, so we fine tune it to 0.0005 in the experiments. Dropout is a technique to prevent overfitting during the training process. Usually, dropout and momentum can improve the learning effect [26]. Since using dropout in all layers will make the network take a long time to converge, the dropout value is set to 0.5, $\alpha = 0.9$, and $\lambda = 0.0005$ in the fully connected layer.

The output of the seventh layer is its input, which contains m neurons corresponding to m types of music styles, and the output probability is $p = [p_1, p_2, \ldots, p_m]^T$, using the softmax regression given in the following equation.

$$p_j = \frac{\exp\left(X_8^j\right)}{\sum_{i=1}^m \exp\left(X_8^i\right)}, \tag{2}$$

where x_8 is the input to the softmax function, j is the current category being computed, and $j = 1, \ldots, m$, p_j represents the true output of the j^{th} category.

4. Experimental Results and Analysis

In this study, the Caffe framework [27] is either the library or interface tool that helps ML developers to develop and design the DL model more professionally. It is used to train CNN models for music style recognition.

4.1. Dataset. The proposed method used a GTZAN style collection database [28] using recognition rate as a performance metric. The GTZAN dataset was collected by [29], which consists of 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock). Each genre category contains 100 audio recordings of up to 30 seconds in length, for a total of 1000 music excerpts. To adjust these

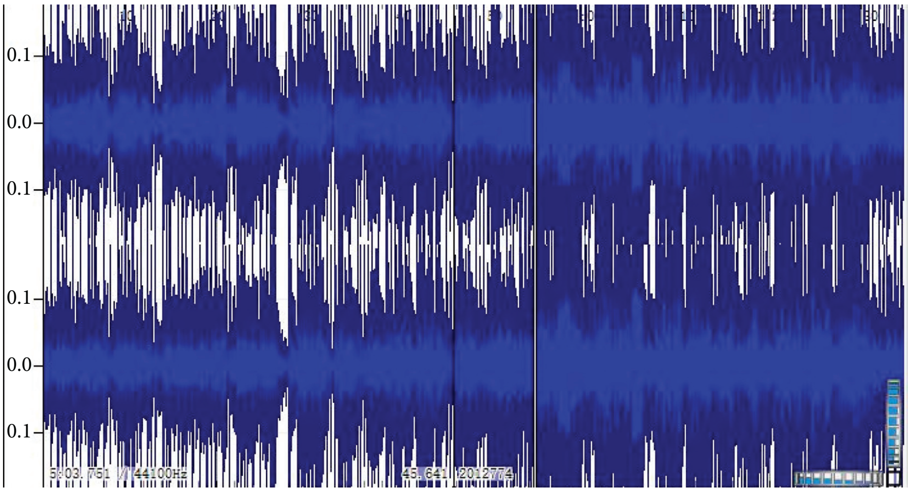

Figure 2: Different spectra after the HPSs algorithm.

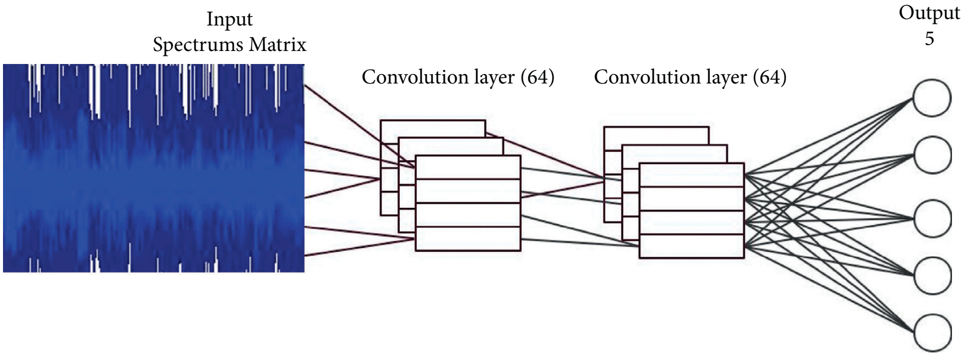

Figure 3: CNN structure of music style recognition.

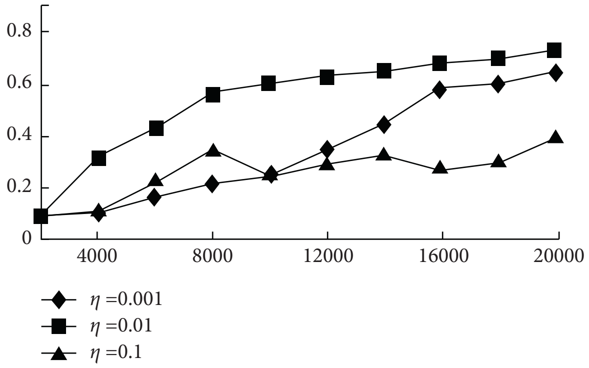

Figure 4: Effect of learning rate η.

hyperparameters, the dataset was randomly divided into two subsets in the ratio of 5 : 1, i.e., 2500 music tracks for training and 500 music tracks for testing.

Figure 4 shows that when the music classification η is relatively small, such as 0.001, the learning process is very slow, and the recognition rate is not yet stable for 20,000 iterations of training samples. The learning efficiency can be improved by increasing η appropriately. At the same time, if η is too large, such as 0.1, the learning process will be unstable and the classification performance will be reduced. Figures 5 and 6 show the effects of momentum μ and weight decay λ. Figure 5 shows that using momentum μ can speed up the learning process well, while if μ is large, such as 0.97, it causes oscillations and slower convergence in the initial stage, and it reduces the classification performance in the later stage. Figure 6 shows the effect of the weight decay λ, showing that a smaller λ seems to be a safer choice, while a larger λ such as 0.005 destabilizes the learning process.

This way, such neurons will not affect both forward and backward propagation, so that for each input sample, a different network structure is used, but the weights are shared, so that the obtained parameters can be adapted to the network structure in different situations, and the generalization ability of the network is improved. In this experiment, the dropout is fine tuned to 0.5 or 0.6. When the dropout value is increased, the training time is slightly longer and the convergence is slower, and the training is carried out for 20,000 iterations. The CNN-based classifier can produce good classification performance after 20,000 iterations. Figure 7 shows the different recognition rates for different dropout values with different number of iterations.

In conclusion, the hyperparameters in the CNN such as music classification η, momentum coefficient μ, weight decay coefficient λ, and dropout values can significantly affect the training process of the network and must be

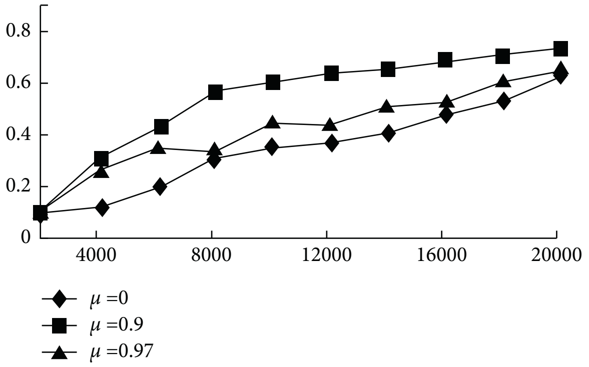

Figure 5: Effect of momentum coefficient.

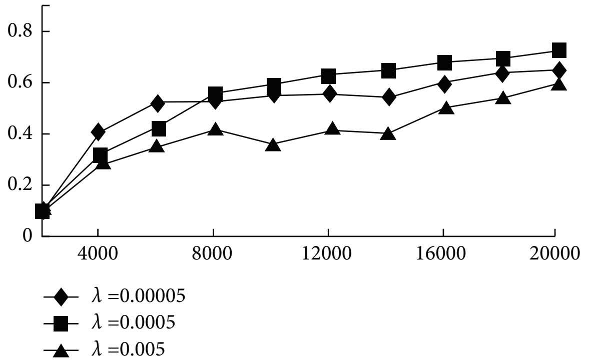

Figure 6: Effect of attenuation coefficient λ.

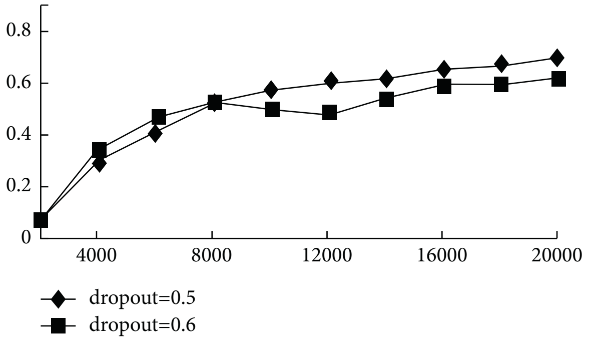

Figure 7: Effect of the dropout value.

carefully adjusted before obtaining satisfactory classification performance. In the experiments of this study, the recognition rate of the GTZAN dataset is about 73% with no data expansion.

5. Conclusion

After entering the digital era, digital music technology has led to the rise of Internet companies. In this study, we use CNNs to extract and classify multiple features in music signals. The network structure of the CNN is designed, and the effect of different parameters in the network structure on the recognition rate is investigated. Based on the experimental results, the proposed work significantly performed well as about 73% on the GTZAN dataset with no data expansion. This will fundamentally change the way music content reaches music users and is a disruptive technology application for the industry [30–32].

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] J. Zeng, X. Zhao, J. Gan, C. Mai, Y. Zhai, and F. Wang, “Deep convolutional neural network used in single sample per person face recognition,” *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 3803627, 11 pages, 2018.

[2] M. Zhao, H. Li, X. Shi, Y. Chan, X. Luo, and T. Li, “Automated recognition of zygote cytoplasmic area (ZCA) in time-lapse imaging (TLI) based on deep convolutional neural network (CNN),” *Fertility and Sterility*, vol. 108, no. 3, p. e239, 2017.

[3] S. Wang, Y. Xing, L. Zhang, H. Gao, and H. Zhang, “Deep convolutional neural network for ulcer recognition in wireless capsule endoscopy: experimental feasibility and optimization,” *Computational and Mathematical Methods in Medicine*, vol. 2019, no. 2, 14 pages, Article ID 7546215, 2019.

[4] D. Bisharad and R. H. Laskar, “Music genre recognition using convolutional recurrent neural network architecture,” *Expert Systems*, vol. 36, no. 4, pp. e12429.1–e12429.13, 2019.

[5] O. Seddati, S. Dupont, and S. Mahmoudi, “DeepSketch: deep convolutional neural networks for sketch recognition and similarity search,” in *Proceedings of the International Workshop on Content-based Multimedia Indexing*, pp. 1–6, IEEE, Prague, Czech Republic, June 2015.

[6] E. Gardini, M. J. Ferrarotti, A. Cavalli, and S. Decherchi, “Using principal paths to walk through music and visual art style spaces induced by convolutional neural networks,” *Cognitive Computation*, vol. 13, no. 2, pp. 570–582, 2021.

[7] Y. K. Yi, Y. Zhang, and J. Myung, “House style recognition using deep convolutional neural network,” *Automation in Construction*, vol. 118, Article ID 103307, 2020.

[8] J. Wang, Y. Li, H. Feng, L. Ren, X. Du, and J. Wu, “Common pests image recognition based on deep convolutional neural network,” *Computers and Electronics in Agriculture*, vol. 179, no. 1, Article ID 105834, 2020.

[9] A. Ahmed, A. Jalal, and K. Kim, “RGB-D Images for object segmentation, localization and recognition in indoor scenes using feature descriptor and Hough voting,” in *Proceedings of the IEEE conference on applied sciences and technology*, Islamabad, Pakistan, January 2020.

[10] A. Jalal, S. Kamal, and D. Kim, “Depth silhouettes context: A new robust feature for human tracking and activity recognition based on embedded HMMs,” in *Proceedings of the 12th IEEE International Conference on Ubiquitous Robots and Ambient Intelligence*, pp. 294–299, Goyangi, Korea (South), October 2015.

[11] A. Jalal, Y.-H. Kim, Y. -J. Kim, S. Kamal, and D. Kim, “Robust human activity recognition from depth video using spatiotemporal multi-fused features,” *Pattern Recognition*, vol. 61, pp. 295–308, 2017.

[12] S. Badar Ud Din Tahir, A. Jalal, and M. Batool, “Wearable sensors for activity analysis using SMO-based random forest over smart home and sports datasets,” in *Proceedings of the IEEE ICACS conference*, Lahore, Pakistan, February 2020.

[13] S. Kamal, A. Jalal, and D. Kim, “Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM,” *Journal of Electrical Engineering and Technology*, pp. 1921–1926, 2016.

[14] S. A. Rizwan, A. Jalal, and K. Kim, “An Accurate Facial expression detector using multi-landmarks selection and local transform features,” in *Proceedings of the IEEE ICACS conference*, Lahore, Pakistan, February 2020.

[15] A. Farooq, A. Jalal, and S. Kamal, “Dense RGB-D Map-based human tracking and activity recognition using skin joints features and self-organizing map,” *KSII Transactions on internet and information systems*, vol. 9, no. 5, pp. 1856–1869, 2015.

[16] S. Kamal and A. Jalal, “A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors,” *Arabian Journal for Science and Engineering*, vol. 41, no. 3, pp. 1043–1051, 2016.

[17] N. I. Yaacob and N. M. Tahir, “Feature selection for gait recognition,” in *Proceedings of the IEEE symposium on Humanities, science and engineering research*, Kuala Lumpur, Malaysia, June 2012.

[18] S. Hizlisoy, S. Yildirim, and Z. Tufekci, “Music emotion recognition using convolutional long short term memory deep neural networks,” *Engineering Science and Technology an International Journal*, vol. 24, no. 3, pp. 760–767, 2020.

[19] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[20] S. Hizlisoy, S. Yildirim, and Z. Tufekci, “Music emotion recognition using convolutional long short term memory deep neural networks,” *Engineering Science and Technology, an International Journal*, vol. 24, 2020.

[21] J. M. Brotzer, E. R. Mosqueda, and K. Gorro, “Predicting emotion in music through audio pattern analysis,” *IOP Conference Series: Materials Science and Engineering*, vol. 482, no. 1, Article ID 012021, 2019.

[22] N. N. Vempala and F. A. Russo, “Predicting emotion from music audio features using neural networks,” in *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, pp. 336–343, Lecture Notes in Computer Science, London, UK, June 2012.

[23] S. Cunningham, R. Harrison, J. Weinel, and R. Picking, “Supervised machine learning for audio emotion recognition,” *Personal and Ubiquitous Computing*, vol. 25, pp. 1–14, 2020.

[24] A. Haruvi, K. Ronen, N. Brande-Eilat, S. Kalev, E. Kay, and D. Furman, “Differences in the effects on human focus of music playlists and personalized soundscapes, as measured by brain signals,” 2021, https://www.biorxiv.org/content/10.1101/2021.04.02.438269v2.full.

[25] W. Yang, Q. Liu, S. Wang et al., “Down image recognition based on deep convolutional neural network,” *Information Processing in Agriculture*, vol. 5, no. 2, pp. 246–252, 2018.

[26] Y. Han and B. W. Hong, “Deep learning based on fourier convolutional neural network incorporating random kernels,” *Electronics, 2021*, vol. 10, no. 16, 2004.

[27] Y. Jia, E. Shelhamer, J. Donahue et al., “Caffe: convolutional architecture for fast feature embedding,” in *Proceedings of the MM 2014-Proceedings of the 2014 ACM Conference on Multimedia; Association for Computing Machinery, Inc*, pp. 675–678, Orlando, FL, USA, November 2014.

[28] S. Shah, R. Mishra, A. Szczurowska, and M Guziński, “Non-invasive multi-channel deep learning convolutional neural networks for localization and classification of common hepatic lesions,” *Polish Journal of Radiology*, vol. 86, pp. e440–e448, 2021.

[29] 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.

[30] M. N. Razali, E. G. Moung, F. Yahya et al., “Indigenous food recognition model based on various convolutional neural network architectures for gastronomic tourism business analytics,” *Information*, vol. 12, no. 322, 2021.

[31] Y. Cai, Y. Song, P. Ni, X. Liu, and X. Li, “Subwavelength ultrasonic imaging using a deep convolutional neural network trained on structural noise,” *Ultrasonics*, vol. 117, no. 7553, Article ID 106552, 2021.

[32] F. Li, M. Liu, Y. Zhao et al., “Feature extraction and classification of heart sound using 1D convolutional neural networks,” *EURASIP Journal on Applied Signal Processing*, vol. 2019, no. 1, pp. 1–11, 2019.