Automatic Music Genre Classification Based on CRNN

Yu-Huei Cheng, *Member, IAENG*, Pang-Ching Chang, Duc-Man Nguyen, and Che-Nan Kuo

***Abstract*—In recent years, machine learning and deep learning technologies are maturing. The Convolutional Neural Networks (CNNs) are applied to all kinds of fields and various CNN-based fusion and combination methods are also appeared one after another. Due to the streaming media rapid growth, therefore the music genre classification is significant in the multimedia world. In order to raise the user's efficiency when searching for different styles of music, we applied CNN combined with Recurrent Neural Network (RNN) architecture to implement a music genre classification model. In the pre-training step, the Mel-Frequency Cepstrum (MFC) is used as feature vector of sound samples. We use Librosa to convert original audio files into their MFC to achieve a sensory pattern close to that of humans hear. In this study, a model is trained by Mel-Frequency Cepstral Coefficients (MFCC) and CRNN method with the accuracy achieve to 43%. This model will continue to be improved in the future to identify the music style by extracting more sound features.**

***Index Terms*—Convolutional Neural Networks; Recurrent Neural Network; music genre classification; Mel-Frequency**

I. Introduction

ARTIFICIAL Intelligence (AI) has been integrated into our lives with the advancement of hardware and the increase of data. For examples, disease assistant diagnosis [1, 2], natural language processing [3, 4], and prediction of stock price [5, 6]. Furthermore, testing techniques are also developed to detect the erroneous behaviors of Artificial Intelligence System (AIS) [7]. This music streaming services are gradually emerging in today's world, many people search the latest popular music through the online music library. However, the huge music library makes it difficult to search for specific genre music. Therefore, a tool that can automatically classify music is an important issue for organizing, searching, retrieving, and recommending music. Due to the difficulty to selecting and extracting appropriate audio features, and the audio data that has appropriate tags is also hard to be gotten. That results in music genre classification is therefore considered to be a challenging task.

In the past, many researchers have devoted themselves to the study of music parameters and proposed methods for classifying music with different genres. Tzanetakis and Cook established the GTZAN data set, whose purpose is to study the classification of machine learning applied to music genres [8]. Scheirer *et al.* described a real-time beat tracking system with audio signals of music. In this system, the filter bank is coupled to a combined filter network that tracks the signal period to produce the main beat and its result [9]. Eve *et al.* used MIDI, pitch and duration as the characteristics of music to classify to obtain good results [10].

In recent years, the methods for applying audio classification by Convolutional Neural Networks (CNNs) are continued to increase. Most of them use log cepstrum or Mel-Frequency Cepstrum Coefficients (MFCC) as input. Log cepstrum is the logarithm operation after the Fourier transform of the signal, and then perform the inverse Fourier transform to obtain the spectrogram. Since the cepstrum was originally used to measure seismic waves, and human perception of sound and seismic waves is similar to the cepstrum, the axis of the spectrogram is logarithmic to conform human perception. Mermelstein [11] is a pioneer dedicated to the development of MFCC. He proved that MFCC in view of the spectrum of human perception audio. Due to the non-linear relationship between the perceived sound level of the human ear and the actual frequency, the MFCC is more in line with the characteristics of the human ear's perceived audio. Bridle *et al.* used Mel-Frequency Cepstral Coefficients (MFCC) and the filter interval obtained by the design is logarithmic [12]. Based on the above viewpoints, this study uses MFCC as the pre-processing format of our experiment to simulate the mode of human ear perception audio.

There are many methods proposed to classify music with different genres. For example, musical instruments and rhythm are used as classification parameters, and pitch, musical instrument, and beat are also used as classification parameters. Both of the above are ways to classify by comparing similarities. Since MFCC is similar to human hearing, the classification of music genres can be achieved by converting audio into their respective MFCC and comparing similarities. Convolutional Neural Networks (CNNs) have excellent effects on data that cannot be rearranged or elements are lost, and Recurrent Neural Networks (RNNs) can arrange the data to be close to the human semantic description. Therefore, we combine the two and take advantages of each to

Manuscript received May 27, 2020; revised January 13, 2021. This work was supported in part by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST108-2221-E-992-031-MY3, MOST108-2821-C-324-001-ES, MOST109-2218-E-005-002, and MOST109-2218-E-005-002.

Y.-H. Cheng is a professor in the Department of Information and Communication Engineering, Chaoyang University of Technology, Taichung 413310, Taiwan (e-mail: yuhuei.cheng@gmail.com).

P.-C. Chang is a master student in the Department of Information and Communication Engineering, Chaoyang University of Technology, Taichung 413310, Taiwan (e-mail: s10830612@gm.cyut.edu.tw)

D.-M. Nguyen is the dean in the International School, Duy Tan University, 254 Nguyen Van Linh, Danang, Vietnam (e-mail: mannd@duytan.edu.vn)

C.-N. Kuo is an assistant professor in Department of Artificial Intelligence, CTBC Financial Management College, Tainan 709, Taiwan (corresponding author. e-mail: fkikimo@hotmail.com).

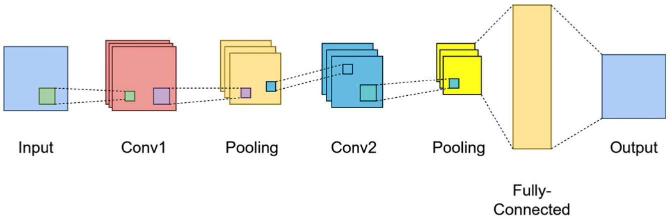

Fig. 1. A CNN model architecture.

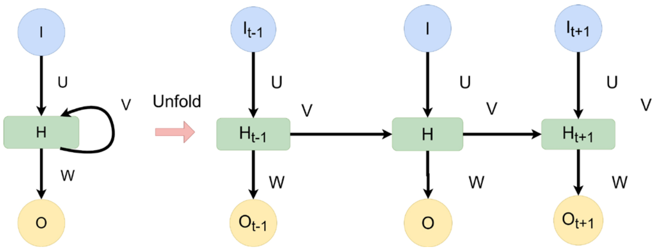

Fig. 2. A RNN model architecture.

build the model.

CNNs have been widely used to solve various complex audio problems. These problems consists sentiment analysis [13], feature extraction [14], genre classification [15], and prediction [16]. The hybrid model of CNNs and RNNs has recently been applied to temporal data such as audio signals and word ordering. Wang and Zhang [17] propose a voice activity detection (VAD) system, which combines the architecture of CRNN and RNN. They found that their architecture has good performance compared with the basic system. Wang *et al.* [18] proposed CRNN-TF to improve the original CRNN architecture and proved that CRNN-TF is superior to CRNN. Wei *et al.* [19] also proposed the application of CRNN in the field of sound event detection (SED). They called the proposed model A-CRNN for adaptive CRNN and applied it to DCASE. The results showed that the architecture can be applied to different data set or different recording equipment.

Therefore, this study is proposed by using a CRNN method for the identification of music genres of different genres.

II. Methods

A. Data Set

GTZAN is a famous data set which was created by Tzanetakis and Cook [8]. The data set contains 1,000 audio tracks, each of which is 30 seconds long. There are 10 genres are included, namely Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, and Rock. Each genre is represented by 100 tracks, and each track is WAV format (*.wav). All are 22,050Hz, monophonic, and 16-bit audio file which has a total capacity of 1.6GB.

TABLE I
PREPROCESSING PARAMETERS

Parameters	Value
Audio length	30s
Sampling rate / sec	660,000
FFT size	2,048
Hop Size	512

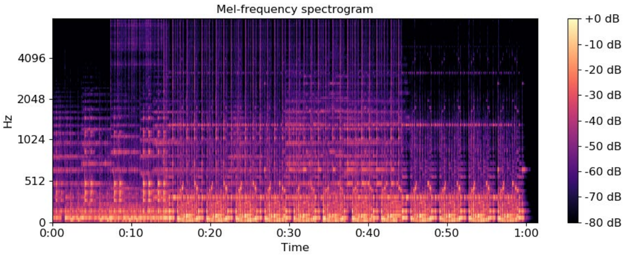

Fig. 3. Mel-frequency spectrogram.

B. Preprocessing

In this study, MFCC was used as input. Aaron *et al.* have used MFCC to pre-process songs [20]. We convert the 30-second WAV audio to AU audio and send it to the pre-processing stage. We use 660,000 sampling rate to quantize the audio signal every second. Also, we perform Fast Fourier Transform (FFT) on 2,048 frames and set the hop size to 512 for the overlapping area between the two frames. The calculation method of the spectrogram is to map the amplitude and frequency of each frame after the fast Fourier transform, and then merge them according to the frame and hop number of the fast Fourier. In order to obtain a logarithmic amplitude spectrogram, the amplitude of the spectrogram is logarithmic. These preprocessing methods can be performed through Librosa. Table I shows the preprocessing parameters.

C. Neural Network Architecture

(1) Convolutional Neural Networks

A CNN consists of one or more convolutional layers, and then one or more fully connected layers. Each layer is composed of multiple neurons. Each neuron calculates the weight after receiving the value from the feature vector, and then transmits the weight to the next layer, as shown in Figure 1.

(2) Recurrent Neural Network

A RNN is a neural network that loops in structure and is commonly used in natural language processing. Since both audio and language are transmitted by waveforms, a RNN can be used to transform data into patterns described by human semantics.

Figure 2 is the RNN model architecture, the left side is the basic model architecture, and the right side is a schematic diagram expanded on the time axis, where I_t is the input at time t, corresponding to H_t, and O_t is the hidden layer and the output layer.

(3) CRNN with fusion of CNN and RNN

This study combines the CNN and RNN methods to propose the CRNN model applying to the classification of different genres of music. The CRNN model used consists of 4 layers of CNN and 1 layer of RNN. The filter we use is a Mel filter built through Librosa. By calculating the MFCC of the audio, the original audio is converted into its own Mel-frequency spectrogram as shown in Figure 3.

The output of the first layer is 32, and it is input to the second layer after passing through the largest pooling layer. The output of the second layer is 64, which is input to the third

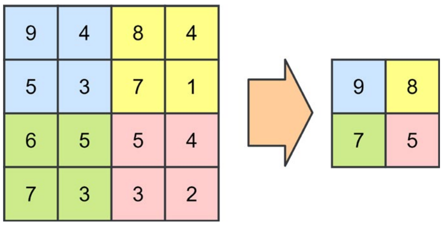

Fig. 4. Schematic diagram of the largest pooling layer.

layer after passing through the largest pooling layer. The output of the third layer is 128, and it is input to the fourth layer after the maximum pooling layer. The fourth layer output is 256, output to the RNN layer. Each layer uses a batch normalization (BN) algorithm. The function of the pooling layer is to reduce the calculation complexity and increase the calculation speed. This study adopts the maximum pooling layer, which is to reduce the matrix by taking the maximum value, as shown in Figure 4.

$$\mu_\beta \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

$$\sigma^2{}_\beta \leftarrow \frac{1}{m}\sum_{i=1}^{m}\left(x_i-\mu_\beta\right)^2 \tag{2}$$

$$\hat{x}_i \leftarrow \frac{x_i-\mu_\beta}{\sqrt{\sigma^2{}_\beta+\epsilon}} \tag{3}$$

$$y_i \leftarrow \gamma\hat{x}_i+\beta \equiv BN_{\gamma\beta}(x_i) \tag{4}$$

In addition, Dropout is also used to reduce the occurrence of overfitting. It is a technique used in deep learning to reduce overfitting. When training a neural network, it is used to randomly disconnect some neurons, that is, these neurons do not participate in the training in the current training. After iterating for optimization, each iteration performs such random sampling to construct a subnet from the original network, and its structure is also different from the original network, so overfitting can be avoided. Figure 5 shows a schematic diagram of Dropout.

Figure 6 is the proposed CRNN model architecture. The size of the input spectrogram is 1,000×1,280×128 (sampling rate×time×frequency). We set the Mel specification to 128×128, and send the audio converted into MFC into the first convolution layer. The size of the first convolutional layer is 32×3×3, and it is sent to the second convolutional layer after the ReLu function, the maximum pooling layer 2×2, and Dropout. The size of the second convolutional layer is set to 64×3×3, and it is sent to the third layer after the ReLu function, maximum pooling layer, and Dropout. The third convolutional layer is set to 128×3×3, and is sent to the fourth convolutional layer after the ReLu function, maximum pooling layer, and Dropout. The fourth convolutional layer is set to 256×3×3, and then sent to the recurrent neural network layer after the ReLu function, the maximum pooling layer, and the Dropout. The recurrent neural network layer is set to 25×128, and finally output after being arranged by the RNN.

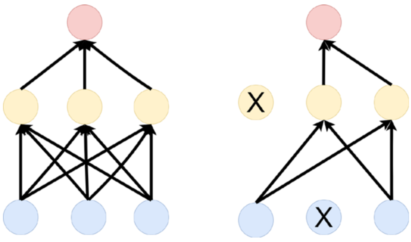

Fig. 5. Schematic diagram of Dropout.

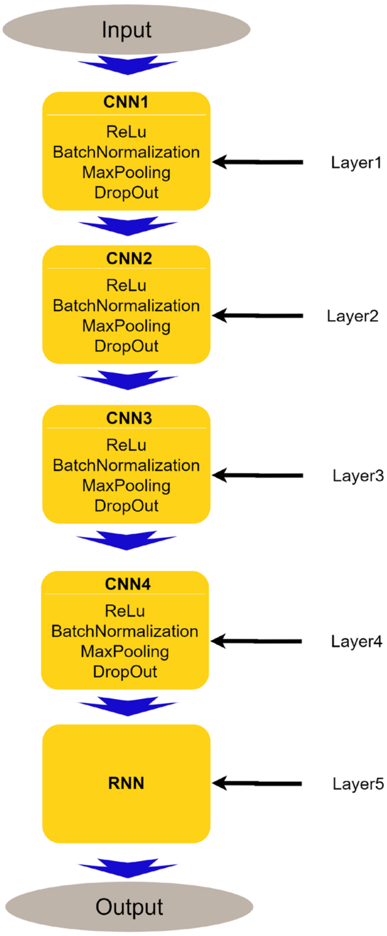

Fig. 6. CRNN model architecture.

During the training process, as the parameters of the previous layer change, the distribution of each layer's input also changes. The internal covariance migration phenomenon requires a lower learning rate to slow down the learning rate, resulting in difficulty in training the neural network. In order to solve the situation of internal covariant migration, BN algorithm is performed after each convolutional layer and before the activation function.

An activation function Leaky ReLu which is a variant of ReLu is added after each layer of convolution, and the function is used. Equation 5 is the mathematical formula of Leaky ReLu, where a_i is a fixed parameter between 1 and +∞.

$$y_i = \begin{cases} X_i, if x_i \geq 0 \\ \frac{x_i}{a_i}, if x_i < 0 \end{cases} \tag{5}$$

TABLE II
COMPARISON OF ACCURACY OF RAW, STFT, AND CRNN METHODS

Methods	Accuracy
Raw	15.00%
STFT	66.00%
CRNN	43.00%

TABLE III
LOSS TABLE

Number of iterations	Loss
0	7
1	5
2	3
3	2.7
4	2.8
5	2.4
6	2
7	2.1
8	1.9
9	1.8
10	1.7
11	1.5

III. EXPERIMENTS AND RESULTS

A. Execution environment

First, we splits the GTZAN data set into 75% training, 10% test, and 15% verification data. Then, we convert the WAV audio file in the data set to the AU audio file and send it to the preprocessing. In the pre-processing stage, all audio is converted into their own MFCC and sent to the model for training. Librosa is a tool for audio signal processing. We use it for pre-processed audio conversion to obtain log-amplitude Mel spectrograms, and then send these spectrograms into our proposed CRNN model for training.

The experiment is to use the Intel i7-6700Q CPU for calculation, equipped with 16GB of memory. We performed 2,400 iterative trainings on the model, the Batch Size was set to 32, the total Epochs was 100, and the operation time was about 70 minutes.

ADAM [21] is suitable for an optimizer that controls the learning rate. It can update the weights of the neural network by the iteration of the training data and perform gradient-based optimization of the objective function. AUC-ROC is a coordinate graph analysis tool, which is a scoring standard commonly used in audio classification, as shown in Equation 6. Among them, true positives (TP) are judged to exist, and actually there are. False positives (FP) are judged to be present, but not actually. A true negative (TN) is judged to be absent, and in fact not. False negative (FN) is judged not to have, but actually there is. AUC-ROC is more suitable for scoring basis of unbalanced data set. Since GTZAN used in this study is a balanced data set (each genre is 100 audio files), we only use accuracy as the scoring standard.

$$ROC = \frac{TP\ /\ (TP + FN)}{FP\ /\ (FP + TN)} \qquad (6)$$

B. Experimental Results

The architecture proposed by this study has an accuracy of 43% in the test data set and a training loss of 0.15. When Epochs is 12 times, the model converges quickly, resulting in reduced accuracy. Since the GTZAN data set is a balanced data set, we use accuracy as a scoring indicator.

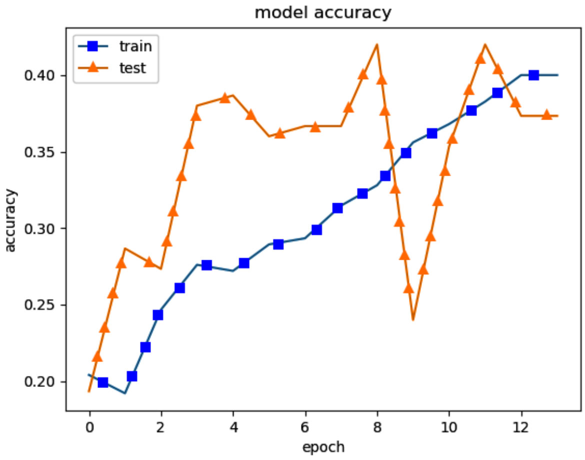

Fig. 7. Test accuracy of the proposed CRNN model.

Table II shows the comparison results of the accuracy of this method and the Raw and STFT [12] methods proposed by Elbir *et al.* Although the CRNN method we proposed is less accurate than the STFT method, the CRNN operation time is shorter. In the future, if the parameters are adjusted, higher accuracy may be achieved. Figure 7 shows the test accuracy of CRNN under different epochs. Figure 8 shows the loss curve of CRNN. Table III shows the unit of loss for each Epochs. We find that the model is close to zero when Epochs is 12, which reduces the accuracy of the model. Therefore, the actual number of iterations is about 350.

IV. CONCLUSION

Music information retrieval can help us understand the context in the audio signal, while music genre classification allows users to have better visibility when selecting their favorite music, and more accurately retrieve and classify different types of music, helping people to reduce search time. In this study, the proposed CRNN model architecture achieves to accuracy rate 43%. Although the accuracy is lower than the STFT method, its training speed is faster. Therefore, this method helps to categorize the huge song database into various genres. In the future, we will focus on

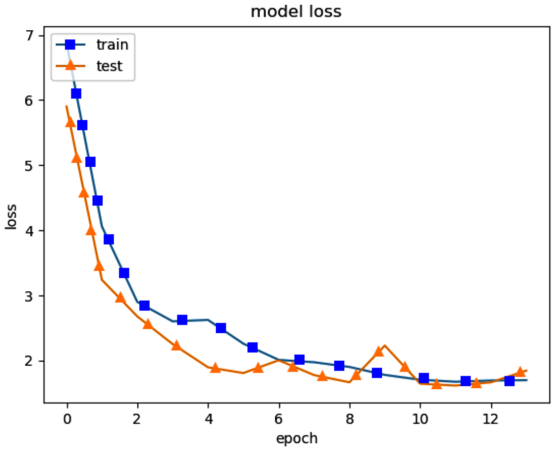

Fig. 8. CRNN model loss curve.

trying to adjust parameters and blend different ways to improve accuracy.

REFERENCES

[1] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine,* vol. 23, no. 1, pp. 89-109, 2001.

[2] D. Li, D. Yang, J. Zhang, and X. Zhang, "AR-ANN: Incorporating Association Rule Mining in Artificial Neural Network for Thyroid Disease Knowledge Discovery and Diagnosis," *IAENG International Journal of Computer Science,* vol. 47, no. 1, pp. 25-36, 2020.

[3] J. Minato, D. B. Bracewell, F. Ren, and S. Kuroiwa, "Japanese Emotion Corpus Analysis and its Use for Automatic Emotion Word Identification," *Engineering Letters,* vol. 16, no. 1, pp. 172-177, 2008.

[4] J. Yan, D. B. Bracewell, F. Ren, and S. Kuroiwa, "The creation of a Chinese emotion ontology based on HowNet," *Engineering Letters,* vol. 16, no. 1, pp. 166-171, 2008.

[5] Z. H. Khan, T. S. Alin, and M. A. Hussain, "Price prediction of share market using artificial neural network (ANN)," *International Journal of Computer Applications,* vol. 22, no. 2, pp. 42-47, 2011.

[6] Q. Zhuge, L. Xu, and G. Zhang, "LSTM Neural Network with Emotional Analysis for prediction of stock price," *Engineering letters,* vol. 25, no. 2, pp. 167-175, 2017.

[7] T. Wu, Y. Dong, Z. Dong, A. Singa, X. Chen, and Y. Zhang, "Testing Artificial Intelligence System Towards Safety and Robustness: State of the Art," *IAENG International Journal of Computer Science,* vol. 47, no. 3, pp. 449-462, 2020.

[8] 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.

[9] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America,* vol. 103, no. 1, pp. 588-601, 1998.

[10] E. Zheng, M. Moh, and T.-S. Moh, "Music genre classification: A n-gram based musicological approach," in *2017 IEEE 7th International Advance Computing Conference (IACC)*, 2017, pp. 671-677: IEEE.

[11] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence,* vol. 116, pp. 374-388, 1976.

[12] J. S. Bridle and M. D. Brown, "An experimental automatic word recognition system," *JSRU Report,* vol. 1003, no. 5, p. 33, 1974.

[13] M. Roopaei, P. Rad, and M. Jamshidi, "Deep learning control for complex and large scale cloud systems," *Intelligent Automation & Soft Computing,* vol. 23, no. 3, pp. 389-391, 2017.

[14] T. L. Li, A. B. Chan, and A. H. Chun, "Automatic musical pattern feature extraction using convolutional neural network," *Genre,* vol. 10, p. 1x1, 2010.

[15] T. Nakashika, C. Garcia, and T. Takiguchi, "Local-feature-map integration using convolutional neural networks for music genre classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[16] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2014, pp. 6959-6963: IEEE.

[17] G.-B. Wang and W.-Q. Zhang, "An RNN and CRNN based approach to robust voice activity detection," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1347-1350: IEEE.

[18] Z. Wang, S. Muknahallipatna, M. Fan, A. Okray, and C. Lan, "Music Classification using an Improved CRNN with Multi-Directional Spatial Dependencies in Both Time and Frequency Dimensions," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1-8: IEEE.

[19] W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-CRNN: A Domain Adaptation Model for Sound Event Detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 276-280: IEEE.

[20] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643-2651.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.

Yu-Huei Cheng (M'12) received the M.S. degree and Ph.D. degree from the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan, in 2006 and 2010, respectively. He crosses several professional fields including biological and medical engineering, electrical and electronic engineering, and information engineering. He is currently a professor of Department of Information and Communication Engineering, Chaoyang University of Technology, Taichung, Taiwan. His research interests include artificial intelligence, automatic control, bioinformatics, biomedical engineering, computational intelligence, embedded systems, electric and hybrid vehicles, internet of things, machine learning, mobile medical, power electronics, renewable energy, and robotics.

Pang-Ching Chang is currently a master student of Department of Information and Communication Engineering, Chaoyang University of Technology, Taichung, Taiwan. His research interests include artificial intelligence, machine learning, and deep learning. Because he interest in the composition of music and the structure of audio, he also covers audio analysis and genre classification, especially the electronic dance music.

Duc-Man Nguyen received his B.S. degree in the Duy Tan University, Danang, Vietnam in 1999, the M.S. degree in the Da Nang University of Technology, Vietnam in 2009, and Ph.D. degree from the Duy Tan University, Da Nang, Vietnam in 2020. Now, he is a Dean of International School, Duy Tan University. His current research interests include software engineering, software testing and automation, software architecture and design, database and database management systems, database integration/ exchange, software/system project management, agile development and testing.

Che-Nan Kuo was born on December 1979 in Tainan, Taiwan. He received his B.S. degree in the Department of Computer Science from the Tunghai University, Taichung, Taiwan in 2002, and the M.S. and Ph.D. degrees from the Department of Computer Science and Information Engineering at the National Cheng Kung University, Tainan, Taiwan in 2004 and 2009. Now, he is an assistant professor in the Department of Artificial Intelligence, CTBC Financial Management College, Tainan, Taiwan. His current research interests include artificial intelligence, interconnection networks, discrete mathematics, computation theory, graph theory, and algorithm analysis.