A deep music genres classification model based on CNN with Squeeze & Excitation Block

Yijie Xu* and Wuneng Zhou†
* College of Information Science, Donghua University, Shanghai 201620, China
† College of Information Science, Donghua University, Shanghai 201620, China
E-mail: zhouwuneng@163.com

***Abstract*—With the development of mobile terminals and Internet technology, people have increasingly convenient mediums to obtain digital music. However, complex music genres and massive music libraries have brought great challenges to music information retrieval. Music genres are high-level labels for music information, which would consume a lot of time and resources when manually tagged. This paper proposes a new model: in order to fully mine the latent information hidden in the input spectrum graph, we build a music genre classification system based on the convolutional neural network that includes Squeeze & Excitation Block (SE-Block), and then use Bayesian optimization to search the best parameters of SE-Block. Finally, we choose the GTZAN dataset for experiments and achieved a classification accuracy of 92%, which is significantly better than most previous research work.**

I. Introduction

Studies have shown that users generally prefer to browse music by genre compared to similarity or recommendation of artists. In addition, genres related to a particular piece of music will also affect preferences [1]. Therefore, most music playback platforms construct music recommendation systems based on music genre classification models, which can provide recommendations to customers or just be launched as commercial products. In this research direction, recognizing genre of music is the first step. It turns out that machine learning technology is very successful in extracting trends and patterns from a large number of databases, and music is inherently suitable for neural network learning thanks to its inseparable relationship with mathematics.

Since deep learning models have proven their superior ability in big data learning in recent years, the use of deep learning in the field of music genre classification has shown a growing trend, and many classification models based on deep neural networks have been proposed [2], [3], [4], [5], [6], [7]. Convolutional neural network (CNN) has shown its excellent performance in many classification tasks in the computer vision field. Since CNN is effective for the classification task of RGB images, the same can be transferred to the classification task of the spectrogram. Medhat et al. [8] proposed a CNN architecture containing masked module. In order to make better use of the acoustic information of music, Costa et al. [9] proposed a classification framework in which both CNN and SVM were involved.

In this paper, we choose the most popular music genre classification dataset GTZAN to train and test our model, but GTZAN always faced problems of insufficient data since its introduction [10]. When the algorithm complexity is high, overfitting is easy to occur. Therefore, this paper has made some improvements to some previous research work. The principal contributions of this paper are as follows.

1) In order to capture as many potential features in the spectrum as possible to improve the performance of the music genre classification model, we built a deep neural network based on CNN and studied the impact of channel number on model performance;
2) We added Squeeze & Excitation Block (SE-Block) to play the role of the attention mechanism in CNN, and used SE-Block to assign weights to the feature map obtained by the convolution operation;
3) At last, we use Bayesian optimization to find the optimal parameter value of the reduction ratio inside the SE-Block.

II. Related Work

A. Convolutional Neural Network in Music Genre Classification Task

In the early stages of CNN's entry into the music genre classification (MGC) field, Sigtia and Dixon [11] verified that ReLU [12], Dropout [13], and Hessian-Free optimizations can improve feature learning effects. In order to take full advantage of the features of CNN's feature extraction, LH et al. [14] first trained CNN as a feature extractor, and then combined the majority voting technology to train feature-based classifiers, so that they obtained significant results on the GTZAN dataset. Unlike the above architecture, Zhang et al. [15] introduced the residual block proposed by He et al. Their model was also obtained good results. In addition, Lin F et al. [16] combines CNN and Bi-direction recurrent neural network, aims to extract spatial information with CNN and temporal information with Bi-direction recurrent neural network; one of the current research trends is the introduction of attention mechanism to learn the potential information of the data, Yang Y et al. [17] introduced three different kind of attention mechanisms in parallel convolutional neural network, the final classification accuracy reached 90%, surpassing all the models mentioned above.

Music genre classification based on CNN is just one of the many branches of music information retrieval. From this point

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on November 29,2025 at 13:25:09 UTC from IEEE Xplore. Restrictions apply.

of view, other tasks of music data can be performed, such as beat tracking, music generation, recommendation system, track separation, and instrument recognition. Music analysis is a diverse and interesting field. Music sessions represent user moments in some way, and finding those moments and describing them is an interesting challenge in the field of data science.

B. Channel Domain's Attention Mechanism

Attention mechanism is a new trend in the research work of music genre classification with neural networks. The basic idea is to let the system learn to pay attention so that it can ignore irrelevant information and focus on important information. Yang Y et al. [17] introduce serial attention and parallel attention mechanisms in CNN respectively, and outperform many previous works.

The focus of the attention mechanism is to assign weights to each feature in the feature vector set. When being input into a CNN, each picture is initially represented by three channels(R, G, B), and then passes through different convolution kernels. After that, each channel will generate a new signal. For example, using a convolutional layer with a kernel-size of 32 for each channel of the image feature will generate a matrix of 32 new channels (H, W, 32), where H and W represent the height and width of the image feature, respectively.

The characteristics of each channel actually represent the components of the image on convolution kernels of different sizes, similar to time-frequency transformation, and the convolution operation of the convolution kernel is similar to the Fourier transform of the signal. Therefore, the information of one channel can be decomposed into signal components on 32 convolution kernels.

Since each signal can be decomposed into components on the kernel function, the new 32 channels generated will definitely contribute more and less to key information. If we add a weight to the signal on each channel to represent the correlation between the channel and the key information, the larger the weight, the higher the correlation, that is, the channel that the network needs to pay attention to.

III. Methodology

As showed in Figure 1, the technical details of each major component of the proposed model are follows.

A. Data Augmentation of Spectrogram

Digital music audio signals can usually be expressed as a function of amplitude and time, with parameters such as frequency, bandwidth, and decibels. Spectrogram is not only the basis for feature extraction in music classification tasks, but also the most intuitive representation of the spectrum when the frequency of a sound changes over time. However, in the MGC task of deep learning, there always exits problems of severe model overfitting and insufficient dataset. Therefore, it is often necessary to perform data augmentation on the spectrogram used for training.

In this paper, the first step in the processing of the spectrogram is audio framing. We use window functions to split different lengths of audio into audio clips of the same size: each track in the dataset is divided by 50% overlaps, the window size which referred to as "n_fft" is 1024 samples, and the distance between the two split windows which referred to as "hop_length" is 512 sampling points. We also tried randomly shifting the signal by scrolling it along the time axis, but the impact on experimental results are not obvious and will not be discussed in detail. After these operations, each 30-second music track is divided into 19 groups of 3-second spectrogram segments, so the total number of spectrogram segments in the dataset becomes 19,000.

B. SE-Block

In CNN, the data transmitted in each convolutional layer exists in a three-dimensional form, which can be treated as multiple two-dimensional pictures superimposed together, each of which is called a feature map.There will be several convolution kernels between the layers. Each feature map of the previous layer and each convolution kernel perform convolution calculations, which will generate a feature map of the next layer. Feature map is an auxiliary tool for CNN to understand pictures, so if a feature map is given weights to make effective feature maps own large weights, and invalid or low-effect feature maps have small weights, the performance of CNN will definitely be improved.

The core idea of the SE-block proposed by [18] is exactly the same. It learns the weight of features through loss. [19] embeds SE-Block in the convolutional layer of CNN, and works on the classification task of multiple time series data sets, and has achieved the state-of-art results.

According to this motivation, the music genre classification model proposed in this paper embeds SE-Block as a substructure in the convolution layer, rather than as a complete structure. This mechanism improves the effects of the original network by assigning weights to the features extracted at each layer, which is equivalent to the attention mechanism for the number of channels in CNN. Although this will inevitably increase the amount of calculations, it does show excellent results in the accuracy of music genre classification.

Figure 2 shows the computation process of SE-Block, F_{tr} represents the convolution operation of the previous layer. The formula is as follows

$$U_c = V_c * X = \sum_{s=1}^{C'} V_c^s * X^s, \tag{1}$$

V_c denotes the c-th convolution kernel, and X^s represents the s-th input. The U output after the F_{tr} operation is the second three-dimensional tensor on the left side in Figure 2, which can also be called as C feature maps of size $H \times W$. And U_c denotes the c-th two-dimensional matrix in U, and the subscript c represents the channel.

F_{tr} is followed by the squeeze operation in SE-Block, and the formula represents a global average pooling

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on November 29,2025 at 13:25:09 UTC from IEEE Xplore. Restrictions apply.

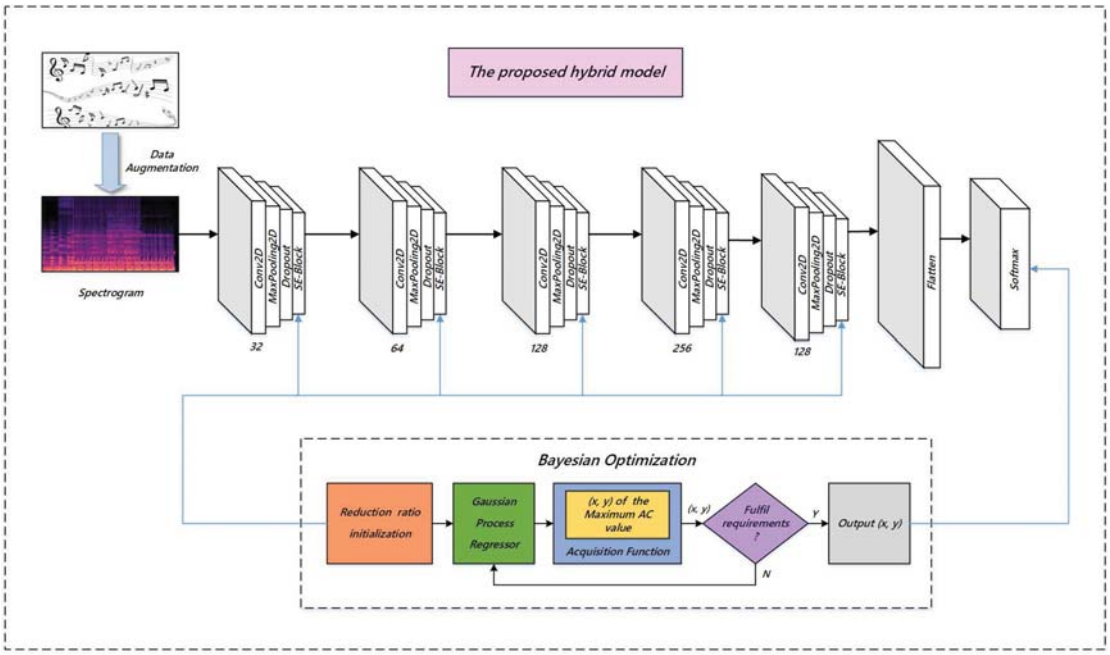

Fig. 1. Framework of the proposed model.

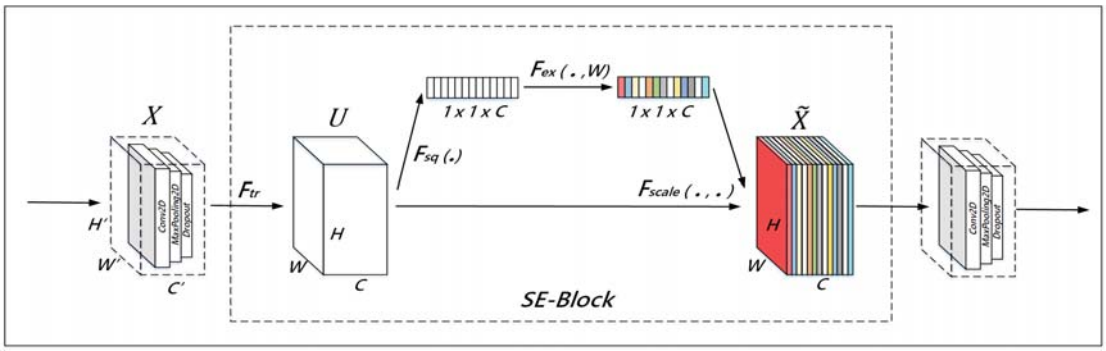

Fig. 2. Computation process of SE-Block.

$$Z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i,j), \qquad (2)$$

Formula (2) converts the input of the $H \times W \times C$ into the output of the $1 \times 1 \times C$. The result of this step is equivalent to indicating the numerical distribution of the C feature maps in this layer, or global information.

The next step is the excitation operation, as shown in Formula (3). The result obtained by squeeze is z. Here we first use a fully connected layer operation to multiply W_1 by z. The dimension of W_1 is $C/r \times C$, and r is a reduction ratio. The purpose of this parameter is to reduce the number of channels so that it can reduce the amount of calculation. And because the dimension of z is $1 \times 1 \times C$, the result of $W_1 z$ is $1 \times 1 \times C$ / r; then it passes through a ReLU layer and the output dimension remains unchanged; then it is multiplied by W_2 which is also a fully connected layer process. The dimension of W_2 is $C \times C/r$, so the dimension of the output is $1 \times 1 \times C$; finally, the sigmoid function is used to obtain s

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \qquad (3)$$

The final s is the core point of SE-Block, the dimension is $1 \times 1 \times C$, and C represents the number of channels. s is used to describe the weight of C feature maps in tensor U, which is denoted the "attention" in SE-Block. And this weight is learned through the previous fully connected layers and non-linear layers. The role of these two fully connected layers is to fuse the feature map information of each channel.

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on November 29,2025 at 13:25:09 UTC from IEEE Xplore. Restrictions apply.

The calculation to be done after getting s is channel-wise multiplication. U_c is a two-dimensional matrix, S_c is a value of weight, so it is equivalent to multiplying each value in the U_c matrix by S_c. Corresponds to F_{scale} in Figure 2, it outputs a $H \times W \times C$ tensor that covers the attention weight of the feature map finally.

$$\tilde{X}_c = F_{scale}(U_c, S_c) = S_c \cdot U_c. \tag{4}$$

C. Bayesian Optimization

Hyperparameters can not be learned from the training process directly, because they are parameters of the algorithm itself. Each model has different hyperparameters, and a good selection of hyperparameters does allow the algorithm to achieve optimal performance. For example, it is necessary to specify the learning rate and the value of reduction ratio *r* in the SE-Block. Manual parameter settings is not only inefficient, but is always affected by human bias, and it is not always possible to find the optimal solution.

Bayesian optimization is a very practical tool for selecting hyperparameters in deep learning researches. It could effectively search for possible hyperparameter spaces and manage a large number of parameters for hyperparameter adjustment. The characteristics in classification models are not suitable for finding hyperparameters with unknown functions and expensive to evaluate, which is where our optimization policy works. The basic method of Bayesian optimization is to estimate the posterior distribution of the objective function from the data, and then select next sample's hyperparameters combination according to the distribution with Bayes' theorem. It takes advantage of the information from previous sample points, then it optimizes by analysing the shape of the objective function and adjusting the parameters that minimize the result to the global minimum [20].

In SE-Block, the reduction ratio *r* denotes the reduced number of channels in each excitation operation, this parameter determines the amount of change in model size in the self-attention mechanism and the number of parameters required to learn these different feature maps. So the application scenario of Bayesian optimization in this paper is

$$r^* = \underset{r \in S}{\arg\min} \ (1 - Accu), \tag{5}$$

in the formula (5), $Accu$ denotes the classification accuracy on testset, S is a candidate set of reduction ratio *r*, and the optimization goal is to select an *r* from S such that the value of $Accu$ will be the largest. The specific formula of model's output is unknown, and is equivalent to a black box function.

Bayesian optimization need to tradeoff the exploration and exploitation for the purpose of avoiding the local optima. Exploitation represents that learning the posterior distribution and then sampling in the areas where the global optimal solution is most likely to occur, and exploration focuses on the areas that have not been sampled [21]. In this paper, our acquisition function chooses expected improvement (EI) which provides a single measure of the usefulness of trying any given point, the computation process are shown in the formula (6) and formula (7)

$$EI(x) = \begin{cases} (\mu(x) - f(x^+)) \cdot \phi(Z) \ + \ \sigma(x) \cdot \phi(Z), & \sigma(x) > 0, \\ 0, & \sigma(x) = 0, \end{cases} \tag{6}$$

$$Z = \frac{\mu(x) - f(x^+)}{\sigma(x)}. \tag{7}$$

$f(x^+)$ is the optimal value we observed in all previous iterations, and ϕ denotes the cumulative distribution function of the standard Gaussian distribution.

IV. Experiments

A. Dataset

In this paper, the GTZAN dataset is selected to train and evaluate our model. The GTZAN dataset was collected by [22] and is widely used in MGC researches. It consists of 1000 audio clips and contains labels for 10 different music genres: Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae and Rock. Each genre contains 100 music clips for 30 s and is stored as a 22,050 Hz, 16-bit mono audio file. This paper divides the training set and test set according to the ratio of 7: 3. After data augmentation process introduced in 3.1, each music sample is divided into 19 groups of spectrogram segments, and the total number of spectrogram segments in the dataset becomes 19,000.

B. Model Specification

The model proposed in our article is implemented using Keras and Tensorflow. During the training phase, we choose the Adam [23] optimizer whose learning rate decreases linearly. The batch size is 64 and we train 50 epochs. We will shuffle all the samples after each period. In addition, a Dropout module is added after each convolutional layer to reduce the trouble of overfitting. The proposed CNN model has five convolutional layers and max-pooling layers. Each layer are added with the SE-Block after the dropout module. The detailed parameters in the CNN with SE-Block are shown in the Table 1.

The number of CNN channels denotes the number of convolution kernels in each convolutional layer. Each channel is an abstraction of the original image. As the number of channels increases, the more information in each layer of the neural network, the less information loss of the original image. In the first three layers of CNN, we want to get more subtle local features from the spectrogram, so choose smaller number of channels, which are 32,64,128, respectively. In the last two layers, we use the number of channels of 256 and 128 to accelerate the reduction of the size of the feature map, gradually abstracting from low-order features to high-order features, and gradually compressing the information. This allows the network to capture more robust features.

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on November 29,2025 at 13:25:09 UTC from IEEE Xplore. Restrictions apply.

As for the SE-Block embedded in each convolutional layer, many related studies ignore the importance of the reduction ratio *r*. The role of *r* is to determine the excitation operation in the SE-Block to reduce the number of channels. The choice of *r*'s size will also affect the performance of the model. However, if the value of *r* is manually selected, there will always be interference from human subjective prejudice, and it will also increase a lot of unnecessary work. So we use Bayesian optimization to select the optimal solution of the reduction ratio *r*. The range of *r* is set to [8, 32], and we use Bayesian optimization for 50 iterations. Finally, according to the result of Bayesian optimization, we set the value of *r* to 31.43, and rounded the result of dividing the channel value by *r* in the excitation operation of SE-Block.

C. Experimental Results

On the GTZAN dataset, the final classification accuracy of our model reached 92%. The optimization of loss function is shown in Figure 3. Based on the same dataset input, the split ratio of the training test set, and the evaluation method, we compare with multiple previous related models in Table 2. From Table 2, we can see that the performance of the proposed model can obtain more accurate classification results.

SE-Block is the core point in our model, while Bayesian optimization policy plays the role to maximize the impact of SE-Block. In experiments, we also found that in music genre classification tasks, changing the number of CNN channels also affected the performance of the model. Therefore, we mainly designed three sets of case studies to explore the influence of these factors on the accuracy of model classification.

1) Case Study 1: Impact of SE-Block: Our original model is a five-layer convolutional neural network. In order to focus on exploring the effects of SE-Block embedded in each layer, we initialize the number of channels of the five convolutional layers to [16,32,64,128,64], The value of *r* is set to 16 first, and the classification accuracy of the original model without adding SE-Block is 83.2. We start with the fifth convolutional layer to embed the SE-Block layer by layer and test the performance of the model. The results are shown in Figure 4.

As we can see from Figure 4, the more layers adding SE-Block, the stronger the classification ability of the model. Compared with embedding the SE-Block in some layers, the experiment shows better results when the module is embedded in all layers. In other words, this can also prove the effectiveness of the channel attention mechanism. The more convolutional layers that include the attention mechanism, the better our network will learn and understand feature maps.

2) Case Study 2: Impact of Channels: When we learned from the last case study that adding SE-Block to each convolutional layer works best, we also need to carefully design the number of channels in the convolutional layer. Because the number of channels has an important relationship with the feature maps extracted by the convolution operation. Generally speaking, the more channels, the more features will be extracted, but too many features sometimes cause overfitting.

TABLE III
DIFFERENT CHANNELS SET'S RESULTS.

Channels parameters	Accuracy(%)
(16, 32, 64, 128, 64)	86.6
(32, 64, 128, 256, 128)	**91.2**
(64, 128, 256, 512, 256)	90.8
(128, 256, 512, 1024, 512)	89.5

After trying four groups of five convolutional layer channels set to [16, 32, 64, 128, 64], [32, 64, 128, 256, 128], [64, 128, 256, 512, 256], [128, 256, 512, 1024, 512], the experimental results are shown in Table 3. After comparison, [32,64,128,256,128] has the best effect, the network can capture and learn enough exquisite features without losing too much information.

3) Case Study 3: Impact of Bayesian Optimiazation: The classification model in different task scenarios has different reduction ratio *r* values when applying SE-Block. The role of *r* is to reduce the feature map whose input is $1 \times 1 \times C$ to the feature map of $1 \times 1 \times C/r$, this step can reduce the amount of calculation later. Generally, when researchers manually adjust the hyperparameters, they will directly select several types of fixed integer values to test the results of the model, as shown in the Table 4:

TABLE IV
MANUALLY HYPERPARAMETERS ADJUSTMENT EXPERIMENTS

Ratio *r*	Accuracy(%)
8	91.0
16	91.2
32	90.8

Using Bayesian optimization to search for hyperparameters in the value range will be more accurate, and can successfully find a model that be able to train with higher accuracy than the results in Table 5. In our experiments, the Bayesian optimization algorithm was instructed to iterate for 50 rounds in the interval [8,32]. We divide the search interval into four smaller intervals, calculate the average and optimal values of the experimental results in each interval, and display them in the table. As we can see from the Table 5, during the optimization process, the better points obtained by the algorithm's search are more concentrated in [8,14.9], but we unexpectedly get the best result at 31.43. If we only use manual adjustment parameters, we cannot get such excellent results.

V. CONCLUSION

In this paper, we have proposed a music genre classification model to implement a small part of the music recommendation system. Based on the spectrogram dataset, this model constructs a convolutional neural network containing SE-Block for effective training. In addition, the hyperparameters searched

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on November 29,2025 at 13:25:09 UTC from IEEE Xplore. Restrictions apply.

TABLE I
MODEL BODY ARCHITECTURE.

Layer	Filter shape	Stride and padding
Conv	$3 \times 3 \times 1 \times 32$	$1 \times 1, 1 \times 1$
MaxPooling	pool: 2×2	$2 \times 2, -$
Dropout	rate: 0.25	-
SE-Block	r: 31.43	-
Conv	$3 \times 3 \times 32 \times 64$	$1 \times 1, 1 \times 1$
MaxPooling	pool: 2×2	$2 \times 2, -$
Dropout	rate: 0.25	-
SE-Block	r: 31.43	-
Conv	$3 \times 3 \times 64 \times 128$	$1 \times 1, 1 \times 1$
MaxPooling	pool: 2×2	$2 \times 2, -$
Dropout	rate: 0.25	-
SE-Block	r: 31.43	-
Conv	$3 \times 3 \times 128 \times 256$	$1 \times 1, 1 \times 1$
MaxPooling	pool: 2×2	$2 \times 2, -$
Dropout	rate: 0.25	-
SE-Block	r: 31.43	-
Conv	$3 \times 3 \times 256 \times 128$	$1 \times 1, 1 \times 1$
MaxPooling	pool: 2×2	$2 \times 2, -$
Dropout	rate: 0.25	-
SE-Block	r: 31.43	-
Flatten	-	-
Dense	Softmax Output:1×10	-

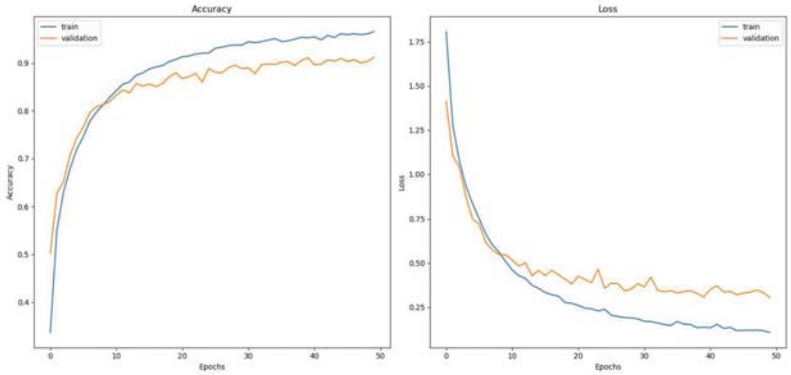

Fig. 3. The classification results obtained by proposed model.

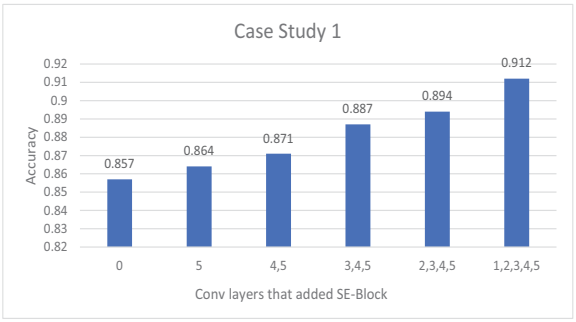

Fig. 4. Results of case study 1.

using Bayesian optimization can enable our model to obtain higher accuracy. Although more experiments are needed to be done, we believe the results are still promising. We have noticed that the channel attention mechanism called SE-Block can effectively enhance the model's understanding of feature maps. In future work, we will study the impact of multiple innovative ensemble learning strategies on model performance. Our next goal is to construct a more user-friendly complete recommendation system. Music genre classification is just one of the many branches of music information retrieval task, and more in-depth research is needed. Our work is still only scratching the surface.

ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (grant no. 61573095). Meanwhile, we are very grateful to editors and reviewers for their insightful comments and suggestions.

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on November 29,2025 at 13:25:09 UTC from IEEE Xplore. Restrictions apply.

TABLE II
EXPERIMENTAL RESULTS ON GTZAN.

Model	Features	Accuracy(%)
CNN + SE-Block + BO	STFT	**92.0**
BRNN + PCNNA	STFT	90.0
nnet1	STFT	84.8
nnet2	STFT	87.4
VGG16	STFT	86.4
KCNN(k = 5) + SVM	Mel-Spectrum, SFM, SCF	83.9
ReLU + SGD + Dropout	FFT (aggregation)	83.0
Multilayer representation	STFT (log representation)	82.0
SVM	STFT	76.2
Random Forest	STFT	70.8
Logistic Regression	STFT	70.0
Decision Tree	STFT	50.2

TABLE V
DIFFERENT *r* INTERVAL'S RESULTS.

Ratio interval	Mean Accuracy(%)	Top Accuracy(%)	Best ratio
[8,14.9]	90.48	91.63	12.69
[14.9, 21.8]	90.39	91.92	19.47
[21.8, 28.7]	90.15	91.14	22.34
[28.7, 35.6]	90.81	**92.02**	**31.43**

REFERENCES

[1] Edward W. Large. Music, thought, and feeling: Understanding the psychology of music. by william forde thompson . new york: Oxford university press, 2008. *Music Perception*, 27(2):145–147, 2009.

[2] Christine Senac, Thomas Pellegrini, Florian Mouret, and Julien Pinquier. Music feature maps with convolutional neural networks for music genre classification. In *the 15th International Workshop*, 2017.

[3] Arjun Raj Rajanna, Kamelia Aryafar, Ali Shokoufandeh, and Raymond Ptucha. Deep neural networks: A case study for music genre classification. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015.

[4] Guangxiao Song, Zhijie Wang, Han Fang, and Shenyi Ding. Transfer learning for music genre classification. In *International Conference on Intelligence Science*, 2017.

[5] Seonhoon Kim, Daesik Kim, and Bongwon Suh. Music genre classification using multimodal deep learning. In *HCI Korea 2016*, 2016.

[6] Nimesh Prabhu, Ashvek Asnodkar, and Rohan Kenkre. Music genre classification using improved artificial neural network with fixed size momentum. *International Journal of Computer Applications*, 101(14):25–30, 2014.

[7] Marco Grimaldi, Pdraig Cunningham, and Anil Kokaram. Discrete wavelet packet transform and ensembles of lazy and eager learners for music genre classification. *Multimedia Systems*, 11(5):422–437, 2006.

[8] Fady Medhat, David Chesmore, and John Robinson. Automatic classification of music genre using masked conditional neural networks. In *2017 IEEE International Conference on Data Mining (ICDM)*, 2017.

[9] Yandre M. G. Costa, Luiz S. Oliveira, and Carlos N. Silla. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied Soft Computing*, 52:28–38, 2017.

[10] Bob L. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *CoRR*, abs/1306.1461, 2013.

[11] Siddharth Sigtia and Simon Dixon. Improved music feature learning with deep neural networks. pages 6959–6963, 2014.

[12] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines vinod nair. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, 2010.

[13] Vladimir Koltchinskii and Dmitry Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics*, 33(4):1455–1496.

[14] Li et al. Automatic musical pattern feature extraction using convolutional neural network. *Lecture Notes in Engineering and Computer Science*, 2180(1), 2010.

[15] Weibin Zhang, Wenkang Lei, Xiangmin Xu, and Xiaofeng Xing. Improved music genre classification with convolutional neural networks. pages 3304–3308, 2016.

[16] Feng et al. Music genre classification with paralleling recurrent convolutional neural network. *ArXiv*, abs/1712.08370, 2017.

[17] Yu et al. Deep attention based music genre classification. *Neurocomputing*, 372:84–91, 2020.

[18] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[19] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. Multivariate lstm-fcns for time series classification. *Neural Networks*, 116, 2018.

[20] Chunfeng Wang, Sanyang Liu, and Mingmin Zhu. Bayesian network learning algorithm based on unconstrained optimization and ant colony optimization. *Journal of Systems Engineering and Electronics*, 23(5):784–790, 2012.

[21] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *NIPS*, 2012.

[22] George Tzanetakis, Student Member, and Perry Cook. Automatic musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[23] L.J. Ba D.P. Kingma. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on November 29,2025 at 13:25:09 UTC from IEEE Xplore. Restrictions apply.