I.J. Information Engineering and Electronic Business, **2020, 1, 30-36**
Published Online February 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2020.01.04

Multi Genre Music Classification and Conversion System

Irfan Siddavatam, Ashwini Dalvi, Dipen Gupta, Zaid Farooqui, Mihir Chouhan
Department of Information Technology, K.J.Somaiya College of Engineering, Mumbai-400077
Email: ashwinidalvi@somaiya.edu

Received: 29 August 2019; Accepted: 25 October 2019; Published: 08 February 2020

Abstract—Artificial Intelligence (AI) has a huge scope in automating, stream- lining, and increasing productivity of Music Industry. Here, we look upon AI based techniques for classifying a piece of music into multiple genres and then later converting it into another user-specified genre. Plenty of work has been done in classification, but using traditional machine learning models which are limited in term of accuracy and rely heavily on features to train the model. The novelty of this work lies in its attempt to covert genre of music from one type to another. This paper focuses on classification achieved by using a model trained via Convolutional Neural Networks. Conversion of music genre, a relatively less worked upon field has been discussed in this paper along with details of implementation. For Conversion, we initially convert the input file to spectrogram. A database of all genre is maintained at all times and a random file from user selected genre is also converted to spectrogram. Later, these spectrograms are processed and converted back to signals. Finally the user can listen to the converted audio file. Validation of the conversion was performed via a survey with the help of end users. Thus, a novel idea of doing Music Genre Conversion was put forth and was validated with positive outcomes.

***Index Terms*—**Music, Artificial Intelligence, Genre, Music Classification, Music Conversion, Convolution Neural Network (CNN)

I. Introduction

THE development of multimedia technology has made digital music widely available to the public. The music files are millions in numbers, making it difficult to identify the genre and segregate them accordingly. Thus, an efficient system should be created to segregate the music files into their respective genres.

Music Genre Classification systems have been a huge part of the music society for over a decade. Genre Classification systems are used to identify the genre of the music file given by the user. Genre is the musical domain to which the song belong viz. Pop Music, Rock Music, Classical Music and so on. Previous work has been done in this field using various approaches, ranging from Machine Learning to Neural Networks. This paper focuses on the approach adopted by us to implement a Multi Genre Music Classification System.

Pushing the idea a bit further, we put forth a novel approach of converting a file from one music genre to another. Aside from being a good exercise in AI, this could be a very good tool to generate new musical ideas.

An extensive literature survey was conducted to gauge the state of the progress made in Music Genre Classification and Conversion. There were many advances made in Classification, chief of which are listed below, and it was established that no work was done in the field of Music Genre Conversion.

As a part of our research, an extensive study was done on multiple papers related to Music Genre Classification. Research done in [3-5] focused on the following genres:

1 Mel Frequency Cepstral Coefficient (MFCC)
2 Short Term Fourier Transform (STFT)
3 Discrete Wavelet Transform (DWT)

These features were considered extremely important to the classification models used in those papers.

The research objective of this work is to apply AI based methods to classify music genre and convert one music genre to another.

Section 2 describes our proposed system for Music Genre Classification and Music Genre Conversion in Section 3 and Sections 4 discuss the results of our endeavours.

II. Proposed Methodology

The proposed classification and conversion system uses GTZAN Dataset which consists of 1000 songs evenly divided in 10 genres. We have selected GTZAN dataset as most of the research papers. The papers we referred during our research mentioned the use of GTZAN and consider it a good, evenly divided dataset of 1000 songs, 30 seconds each [19].

The genres the system currently works with are:

1. Blues
2. Classical
3. Country
4. Disco
5. Hiphop

6. Jazz
7. Metal
8. Pop
9. Reggae
10. Rock

The above genres are available in GTZAN dataset for conversion.

A. Implementation of Classification System

Like most of the tasks, audio processing consists of 3 important tasks. They are:

1. Pre-processing
2. Training and Validation
3. Testing

These tasks have to be executed one after the other in the written order. No task can be skipped as each is important in its own way.

1. Pre-Processing

The objective of this task is to extract relevant audio information. It helps to reduce the dimensionality of the data. It simplifies the learning process. It is important to get this task right as it influences the learning model greatly. Audio pre-processing must be able to extract data from the music. The goal of this task is to extract and organize data in such a manner that it helps the model perform well.

The implemented pre-processing includes the following tasks:

a Creation of windows
b Data Extraction

a Creation of Windows

In this task the audio is divided into windows. This helps in reducing the dimensionality of the data to be extracted. Every window is labelled as the genre it belongs to. These windows will then be used as input to the model after further processing. The creation of windows is done such that it overlaps a little from the previous window which will help the learning model immensely.

b Data Extraction

Data extraction is performed to extract relevant data from the given audio file. For this, we use Librosa, a python package for music and audio analysis. Librosa can be used for extraction of multiple features. We are interested in extracting the entire data from the file, thus we extract spectrograms.

2. Training and Validation

This section discusses the training and validation activities used in implementation of the system.

a Training

An important step in development of an accurate and well performing model is training the model. There are various parameters that greatly influence the output and accuracy of the model. The parameters like batch size, epochs, optimizer, loss, etc have effect on model output. The training task is the most important task in creating a machine learning model. The objective of this task is to generate a model that produces a good accuracy that can be applied for general unseen data. For the training technique we have implemented in this paper we have used Adam optimizer, number of epochs are 50, the batch size is 32 and for loss we are considering categorical-cross entropy.

b Model

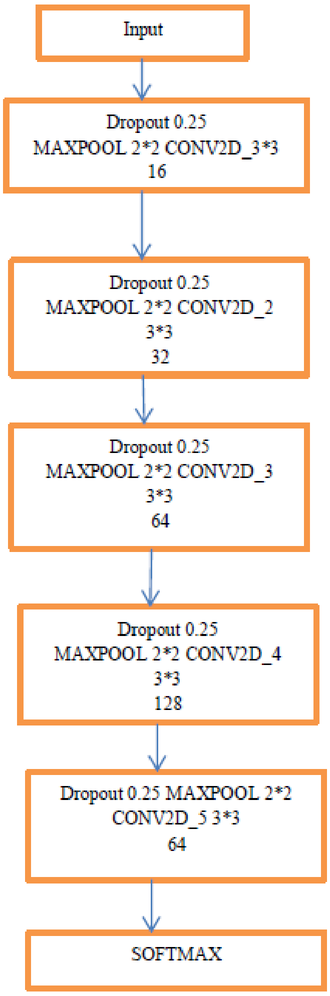


Fig.1. Classification Model

The model has been created using keras as follows: The model that we have implemented in this paper consists of convolutional 2D layers to extract features and then softmax layer to predict the classification of genre.

The figures from [2 – 7] includes discussion on model. The model has an input of size 13300x128x129x1 meaning there are 13300 data points of size 128x129 with a single color channel. The first layer of the model is a convolutional 2D layer of kernel size 3x3 and 16 filters.

The second layer is maxpooling layer with 2x2. The next is a dropout layer of 0.25 which is used to avoid overfitting. These 3 layers together can be said to make one block. Similarly, there are four more blocks with filters 32, 64, 128 and 64 respectively. However, in the 5th block, we have the maxpooling layer with 4x4. The layer used next is Flatten which basically converts the entire output in a row vector. The last layer is a dense layer with softmax activation to perform learning from features extracted from above CNN. The default value for learning rate of Adam optimizer has been used in the model.

The first layer retains the full shape of the spectrogram generated. The activations in the first layer retain almost all of the information in the input picture. As we go deep into the layers, the activations are abstract and less visually interpretable. These layers carry less information about how data has been visually presented and carry information more related to the class.

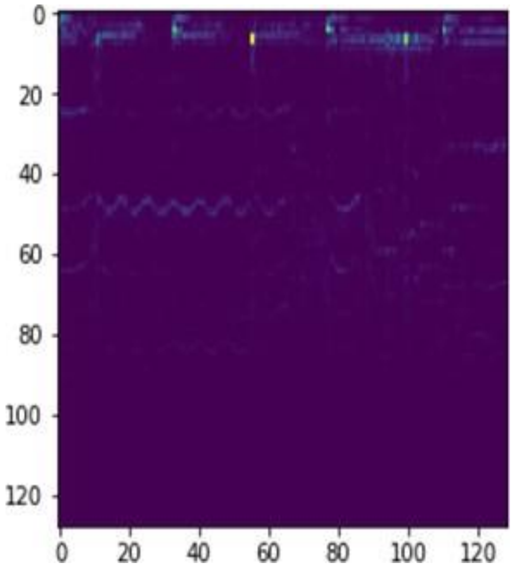


Fig. 2. Sample input for the model

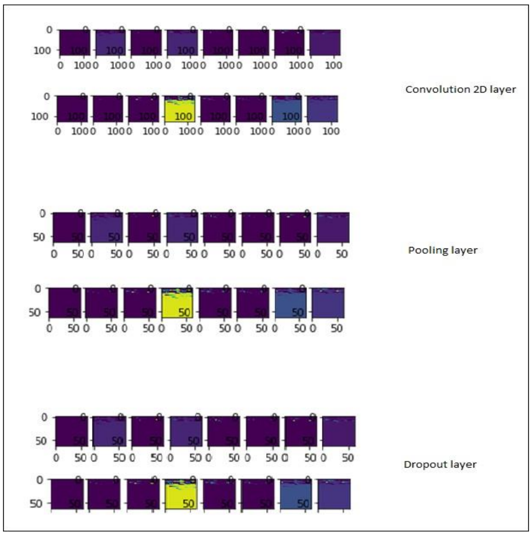


Fig. 3. Convolution Block 1

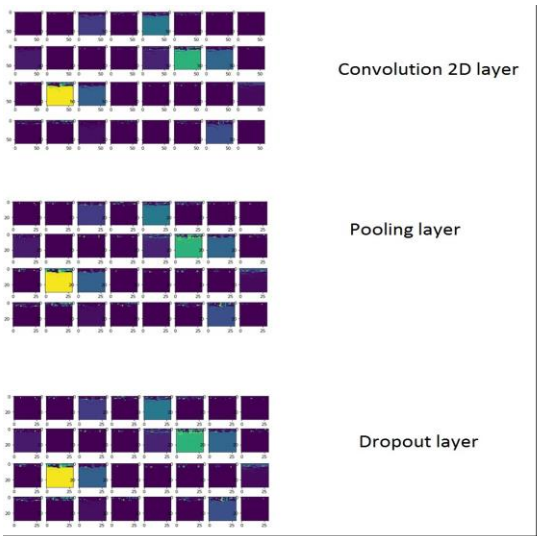


Fig. 4. Convolution Block 2

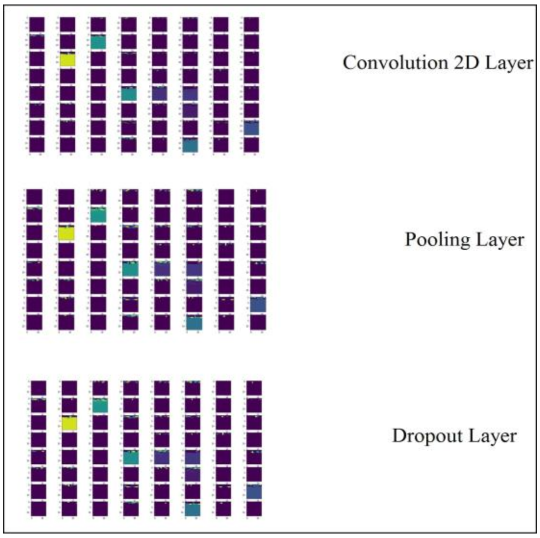


Fig. 5. Convolution Block 3

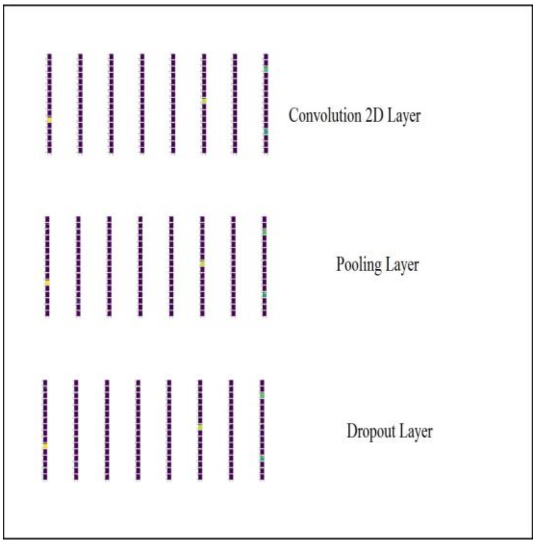


Fig. 6. Convolution Block 4

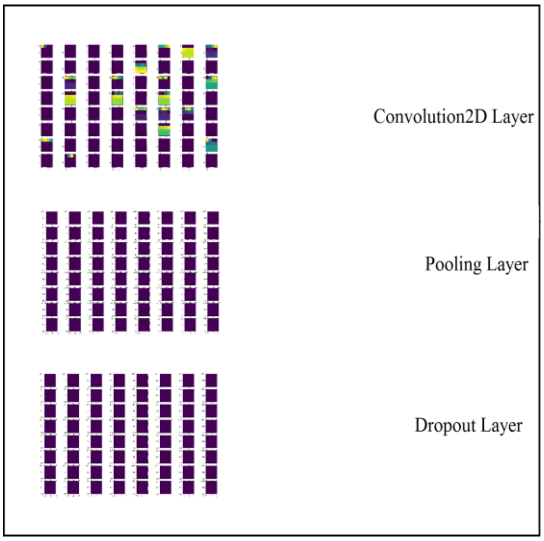


Fig. 7 Convolution Block 5

c Validation

Validation is the process of evaluating the accuracy of the model. The testing dataset is a separate portion of the same data set from which the training set is derived. The dataset used for the system consisted of 1000 songs. The training- testing split was 70 percent and 30 percent. As previously mentioned, the song was divided into multiple windows which were used for training and testing. Each song had 19 windows, a total of 19000 windows for the entire dataset. So, in total 13,300 windows were used for training the network and 5700 for validation purposes. The accuracy for the model was 83.37 percent.

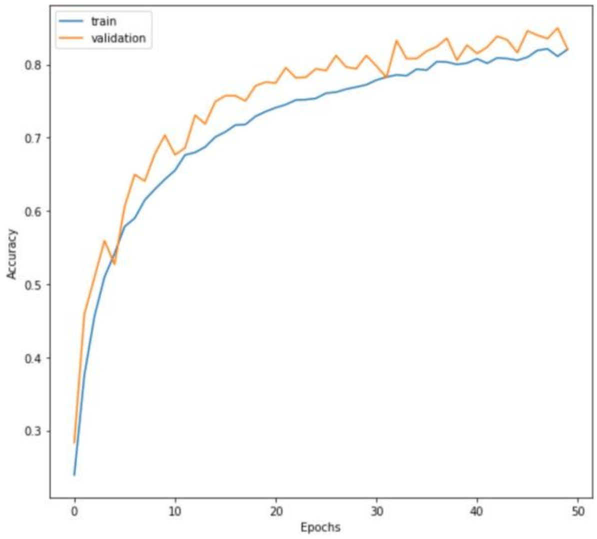


Fig. 8. Accuracy over iterations

3. *Testing*

Testing is the process where one sees if the model created runs properly for unseen real data. After training was completed, model was able to accept the testing audio file, dividing it into 19 windows. Each window is labelled which enables us to derive number of conclusions for the genre of the song as all 19 windows collectively make one song. We can go for multi genre classification or just classify it into a single genre.

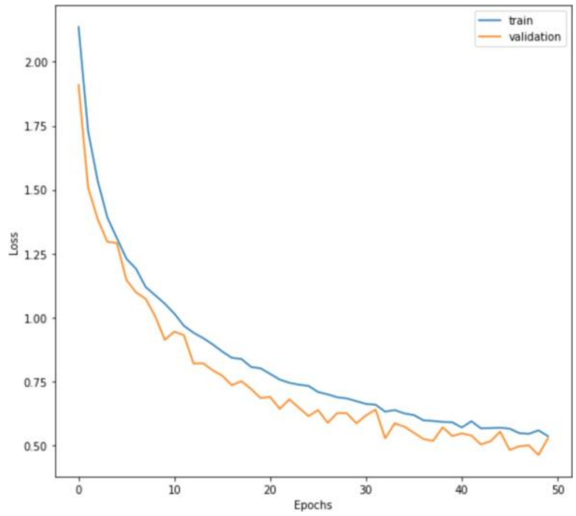
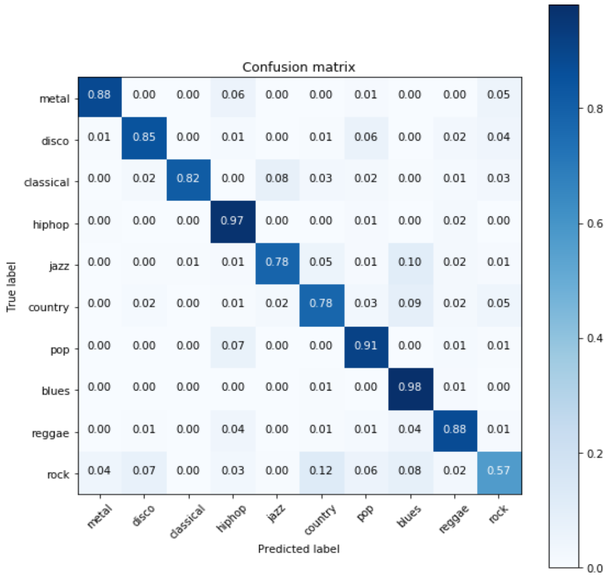


Fig. 9. Loss over iterations



	metal	disco	classical	hiphop	jazz	country	pop	blues	reggae	rock
metal	0.88	0.00	0.00	0.06	0.00	0.00	0.01	0.00	0.00	0.05
disco	0.01	0.85	0.00	0.01	0.00	0.01	0.06	0.00	0.02	0.04
classical	0.00	0.02	0.82	0.00	0.08	0.03	0.02	0.00	0.01	0.03
hiphop	0.00	0.00	0.00	0.97	0.00	0.00	0.01	0.00	0.02	0.00
jazz	0.00	0.00	0.01	0.01	0.78	0.05	0.01	0.10	0.02	0.01
country	0.00	0.02	0.00	0.01	0.02	0.78	0.03	0.09	0.02	0.05
pop	0.00	0.00	0.00	0.07	0.00	0.00	0.91	0.00	0.01	0.01
blues	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.98	0.01	0.00
reggae	0.00	0.01	0.00	0.04	0.00	0.01	0.01	0.04	0.88	0.01
rock	0.04	0.07	0.00	0.03	0.00	0.12	0.06	0.08	0.02	0.57

Fig. 10. Confusion Matrix

B. *Implementation of Conversion System*

The following is a high-level block diagram of the conversion system:

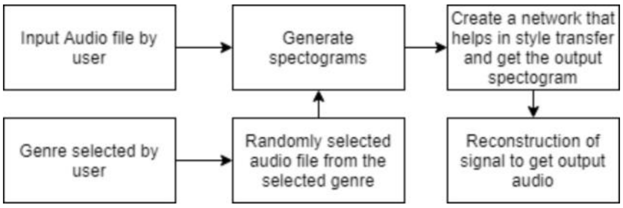

Input Audio file by user
Generate spectograms
Create a network that helps in style transfer and get the output spectogram
Genre selected by user
Randomly selected audio file from the selected genre
Reconstruction of signal to get output audio

Fig. 11. Implementation of conversion system

Our approach to this problem is rather simple. We use the concept of style transfer used in image processing in our conversion module. The steps are:

1. Data Extraction
2. Creation of Network
3. Reconstruction of Signal

1. Data Extraction

Convert the input file to its spectrogram using STFT (Short Time Fourier Transform). Spectogram being a 2D representation of a 1D signal can be thought of as 1xT image with F channels. Since, we are not synthesizing music we take 2 inputs from the user, the audio file and the genre the user wants it to be converted to. We have a database of all the genres and we randomly select a music file from the genre that user wants his file to be converted to. Both files are then converted to its spectrogram using STFT

2. Creation of Network

We create a network using random weights. The network has only one layer with 4096 filters. We first compute features of the spectrograms we get from the previous step using convolution 2d network. We then minimize the loss using Tensorflows ScipyOptimizerInterface using L-BFGS-B method and maximum iterations set at 300.

3. Reconstruction of Signal

From spectrogram to signal, the inversion is done using librosa libraries. Article[18] is the inspiration for algorithm that is used for generating signal from spectrogram. This signal generation helps us to get the output as an audio signal which we can listen to.

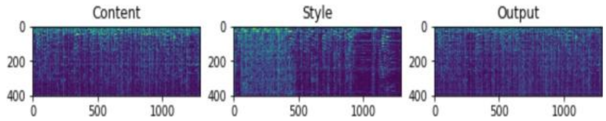


Fig. 12 Here the first image is spectrogram of users input. The second image is spectrogram of the audio which the users input has to take style of. The third image is the result of style transfer from music of style to content.

III. Result

From our implementation and previous research, we can conclude that Multi Genre Classification model using Convolution Neural Networks in which individual windows are used for training and testing instead of features extracted from audio files, gives a better result in terms of accuracy compared to the traditional machine learning models. Moreover, the classification is multi genre where a single audio file can be classified into multiple genres at the same time. Accuracy obtained from this process was 83.37 percent. Graphs and Confusion Matrix obtained as a part of the process have been included in this paper. Moreover, random files selected for testing were classified to their respective genres correctly.

The result is further verified by survey method.

Almost everybody has their own different perspective about music, and what makes one genre of music different from the other. Thus, for qualitative analysis we decided to get user reviews and their feedback to rate the system. We created a sample web page where users can listen to the original file and the converted file for each genre.

A Google Form was embedded in the page, which collected the following information:

1. Name
2. Age
3. Gender
4. Music Knowledge
5. Rating of the system
6. Comments, if any

These attributes were collected for the following reasons:

1. Age: Different age groups have a different music perspective. Thus, what seems likable to people of a certain age group might not seem the same to a different age group. For example, people above 40 years of age will prefer old or classical songs, whereas people belonging to the 15-25 years old category will prefer pop or electronic songs.
2. Gender: People belonging to different gender have a different taste in music. Article [17] states women prefer the piano, violin, and the flute while men prefer guitars, drums and the trumpet. Of course this is one aspect of a study and subject to debates.
3. Music Knowledge: People who have been involved in music for some time and who have a decent knowledge of various genres and aspects related to music will be in a better position to judge the converted audio file. Such people can contribute to the project with constructive feedback.
4. Rating of the system: Finally, we asked users for their ratings on the system. Average ratings were considered while creating this result.
5. Comments: We asked people for their comments on the system, so we can understand the areas of improvement from user perspective.

A total of 46 people gave their reviews about our system. To simplify the results, we divide the user base into 3 groups. Users with Less experience in music (those who rated themselves 1 or 2 on a 5 point scale), users with Decent experience in music (rated 3 on a 5 point scale) and Good experience in music (4 or 5 on a 5 point scale).

Following results were noted:

Table 1. Result of survey

Group	Average Age	Rating of Conversion System
Less Music Experience (1/5 or 2/5)	18-21	3.5/5
Decent Music Experience (3/5)	18-21	3.75/5
Good Music Experience (4/5 or 5/5)	18-21	4/5

From results we can conclude that people with good experience in music found the system satisfactory compared to people with less experience. Moreover, feedback obtained from users ranged from appreciation to comments about improving the output of the system. These comments provided us insights on the notable features of the system and areas of improvements.

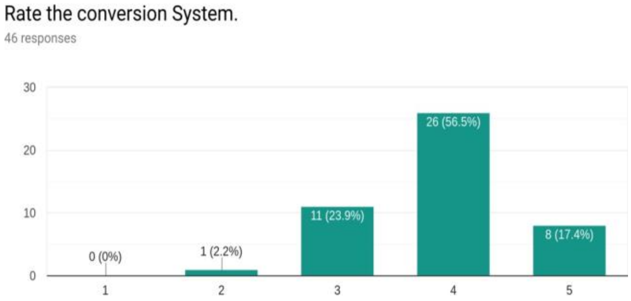


Fig.13. Rating by Users

IV. Conclusion

Music Genre Conversion, a rather less worked and implemented field was discussed and method of implementation was introduced. It was noted that the use of spectrogram via STFT to convert the user given input file and another randomly selected file belonging to the genre of user's choice provided a better result in terms of conversion.

A formal survey of the conversion system was conducted where users were asked to review the system. The results noted and discussed in the paper.

There is a lot of scope for development is the project. The alpha value mentioned in the implementation of the conversion section is currently static which limits the conversion system and some songs are not readily understood. As a part of our future work, we need to make the alpha value dynamic and conduct an extensive study of every genre to understand what alpha values suits what genre best.

References

[1] Pacha, Alexander, and Horst Eidenberger.(2017) "Towards Self-Learning Optical Mu- sic Recognition." Proceedings of the 16th IEEE International Conference On Machine Learning and Applications.

[2] Reed, Jeremy, and Chin-Hui Lee. (2011) "Preference music ratings prediction using tokenization and minimum classification error training." IEEE Transactions on Audio, Speech, and Language Processing 19.8: 2294-2303.

[3] Tsunoo, Emiru, et al. (2011) "Beyond timbral statistics: Improving music classification using percussive patterns and bass lines." IEEE Transactions on Audio, Speech, and Language Processing 19.4: 1003-1014.

[4] Lu, Jing, et al. (2009) "Music style classification using support vector machine.", 452-455.

[5] Ren, Jia-Min, Ming-Ju Wu, and Jyh-Shing Roger Jang. (2015) "Automatic music mood classification based on timbre and modulation features." IEEE Transactions on Affective Computing 6.3: 236-246.

[6] Ridoean, Johanes Andre, et al. (2017) "Music mood classification using audio power and audio harmonicity based on MPEG-7 audio features and Support Vector Machine." Science in Information Technology (ICSITech), 2017 3rd International Conference on. IEEE.

[7] Quinto, Rene Josiah M., Rowel O. Atienza, and Nestor Michael C. Tiglao. (2017) "Jazz music sub-genre classification using deep learning." Region 10 Conference, TENCON 2017-2017 IEEE.

[8] Srinivas, M., Debaditya Roy, and C. Krishna Mohan. (2014) "Music genre classification using On-line Dictionary Learning" Neural Networks (IJCNN), 2014 International Joint Conference on. IEEE.

[9] Ren, Jia-Min, Ming-Ju Wu, and Jyh-Shing Roger Jang. (2015) "Automatic music mood classification based on timbre and modulation features." IEEE Transactions on Affective Computing 6.3: 236-246.

[10] Xue, Angela, and Nick Dupoux. "Predicting A Songs Commercial Success Based on Lyrics and Other Metrics."

[11] Fu, Zhouyu, et al. (2011) "A survey of audio-based music classification and annota- tion." IEEE transactions on multimedia 13.2: 303-319.

[12] Karatana, Ali, and Oktay Yildiz. (2017) "Music genre classification with machine learning techniques" Signal Processing and Communications Applications Conference (SIU), 2017 25th. IEEE.

[13] Panagakis, Yannis, Constantine L. Kotropoulos, and Gonzalo R. Arce. (2014) "Music genre classification via joint sparse low-rank representation of audio features." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 22.12: 1905-1917.

[14] Kravitz, Aaron, Eliza Lupone, and Ryan Diaz. "From Classical To Hip- Hop: Can Machines Learn Genres?." folk 13.192: 22-13.

[15] Benetos, Emmanouil, and Constantine Kotropoulos. (2010) "Non-negative tensor fac- torization applied to music genre classification." IEEE Transactions on Audio, Speech, and Language Processing 18.8: 1955-1967.

[16] Su, Shih-Yang, et al. (2017) "Automatic conversion of Pop music into chiptunes for 8- bit pixel art." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE.

[17] Musical Taste Differences Between Men And Women - https://joyruffen.com/musical-taste-differences-men-women/

[18] Audio Texture and style transfer - https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/

[19] http://marsyas.info/downloads/datasets.html

Authors' Profiles

Ashwini Dalvi joined the Department of Information Technology, K.J.Somaiya College of Engineering, Mumbai in 2006 as an Assistant Professor. She has published over 25 journal and conference papers in the areas of Security, Intelligent applications.

Irfan Siddavatam has received the PH.D. Degree from VJTI, affiliated to Mumbai University, in 2018.

In 2001, he joined the Department of Information Technology, K.J.Somaiya College of Engineering, Mumbai, as an Associate Professor. His research interests include Cyber Physical System Security, Artificial Intelligence, and Internet of Things.

How to cite this paper: Irfan Siddavatam, Ashwini Dalvi, Dipen Gupta, Zaid Farooqui, Mihir Chouhan, " Multi Genre Music Classification and Conversion System", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.12, No.1, pp. 30-36, 2020. DOI: 10.5815/ijieeb.2020.01.04