International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 06th Feb 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=ISSUE-02

DOI: 10.48047/IJIEMR/V12/ISSUE 02/23

Title **Music Genre Classification using Spectrogram Imagesand Tabular Data with Extracted Features**

Volume 12, Issue 02, Pages: 144-149

Paper Authors

Prof. Beebi Naseeba, Mr. Niranjan S Nair, Mr. S Bhaskar Nikhil, Mr. Arjun P,Mr. Akash RJ, Mr.Panyam Venkatesh

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Music Genre Classification using Spectrogram Imagesand Tabular Data with Extracted Features

[1]Prof. Beebi Naseeba, [2]Mr. Niranjan S Nair, [3]Mr. S Bhaskar Nikhil, [4]Mr. Arjun P, [5]Mr. Akash RJ, [6]Mr.Panyam Venkatesh

[1]School of Computer Science and Engineering (SCOPE), VIT - AP University
Amaravati, India, beebi.naseeba@vitap.ac.in

[2]School of Computer Science and Engineering (SCOPE), VIT - AP University
Amaravati, India, niranjansnair18@gmail.com

[3]School of Computer Science and Engineering (SCOPE), VIT - AP University
Amaravati, India, sbnikhil2002@gmail.com

[4]School of Computer Science and Engineering (SCOPE), VIT University, Chennai, India
arjun.purushothaman2020@vitstudent.ac.in

[5]School of Computer Science and Engineering (SCOPE), VIT University, Vellore, India
akash.rj2020@vitstudent.ac.in

[6]School of Computer Science and Engineering (SCOPE), VIT - AP University
Amaravati, India, panyamvenkatesh99@gmail.com

Abstract

Music genres are essentially tags or labels assigned by humans to simplify the identification of the nature of different types of music. They are generally characterized by the overlapping characteristics and features common to music belonging to that particular genre. These properties are usually associated with the instrumentation and rhythmic structures such as the waves in a musical piece. As the annotation or assignment of musical genres is performed manually, it is essential for automation. Automatic music genre classification using deep learning can make it an easier job and helps in improving user experience in apps. In this paper, music is classified into 10 genres from the GTZAN dataset using certain audio features that we explore in depth. Two different types of data are used to classify the music, which are tabular using ANNs and spectrogram images using CNN transfer learning. For the ANN task an accuracy of 95% was procured and for the tabular data with already extracted features an accuracy of 80% was obtained for the CNN task for images of spectrogram.

Keywords – Genre Classification, Transfer Learning, Convolution Neural Networks, Mel Spectrogram images, ANN, Short Time Fourier Transform (STFT)

Introduction

Music genre refers to a style or type of music. Often people have preferences for what kind of music they like and what they would like to listen to at that moment. Popular music streaming platforms like Spotify, Sound Cloud and iTunes want to recommend music of some genre to the people who frequently listen to that genre. This is made possible using music genre classification models that can help to classify music according to various audio features or images. The most essential and basic step in the task of automatic music recommendation is labelling and classifying music by its genre. Most of the current music genre classification techniques use machine learning techniques. We can improve the quality and complexity of the predictions and models using deep learning techniques and neural networks. We perform two different tasks, that is, classification using a tabular dataset containing the extracted features of the audio files and Mel spectrogram images. Generally, in deep learning we use artificial neural networks or ANNs for working through tabular data and convolutional neural networks for CNNs for image data. The image data consists of Mel spectrograms. First, we try to understand what a spectrogram of an audio file is:

Spectrogram: In simple words, A spectrogram is a graphic that shows the frequency spectrum of an audio recording over time. This indicates that as the colours in the figure increase brighter, the sound is more densely concentrated around those particular frequencies, and as the colours get darker, the sound gets closer to being empty or dead sound.

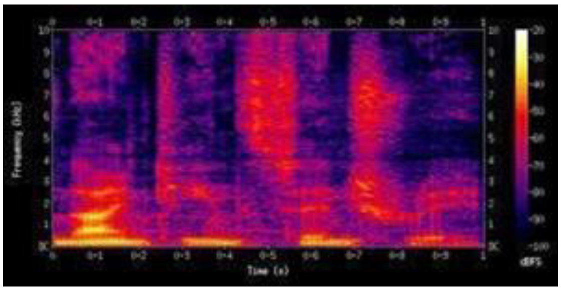


Fig 1: Digitally produced spectrogram

Source: Adapted from [4]

Mel spectrogram: Spectrograms that have been converted to the Mel scale are known as Mel spectrograms. It is a representation of the frequency spectrum of a signal, where a signal's frequency spectrum is its range of frequencies. A perceptual scale of pitches that listeners perceive to be equally spaced from the other is known as the Mel scale (named after the word melody).

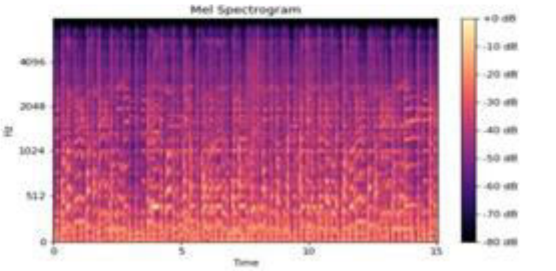


Fig 2: A sample Mel spectrogram

Convolutional neural networks commonly called CNNs, have been widely and profitably applied to deep learning in recent years for a variety of image categorization tasks. While this is going on, Sander et al. [13] demonstrate that spectrograms of music audio can also perform well with CNNs when compared to standard images. In this situation, there is an increasing propensity to train CNNs to learn robust feature representations from music spectrograms. In the upcoming sections we explore the methods we used for data processing and the ANN and CNNmodels used.

Literature Review

The paper referenced for this model [1] uses a custom-made sequential model with convolution and max pooling layers which is preceded by feature extraction and data modification. With their model they got an accuracy of 73%. As the amount ofimage data is not large, we do not get an extremely high accuracy for the CNN models. For now, we work only on the existing data. We managed to improve this accuracy using transfer learning method instead of creating a model from scratch.

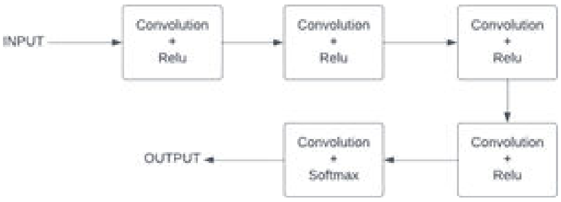
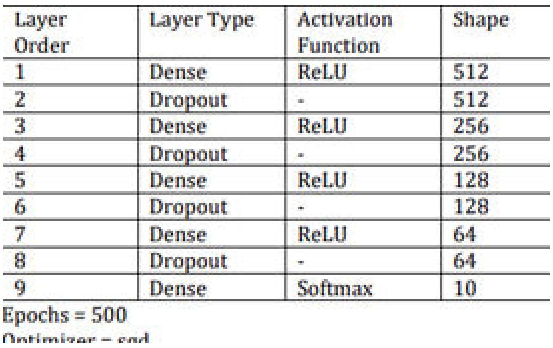


Fig 3: Architecture of existing CNN model

Source: Adapted from [1]

For the ANN task, the referenced paper [3] has used multiple ANNs to perform the task on the same dataset that we used. Their best model got an accuracy of 91% on the test data without overfitting too much.



Layer Order	Layer Type	Activation Function	Shape
1	Dense	ReLU	512
2	Dropout	-	512
3	Dense	ReLU	256
4	Dropout	-	256
5	Dense	ReLU	128
6	Dropout	-	128
7	Dense	ReLU	64
8	Dropout	-	64
9	Dense	Softmax	10

Epochs = 500
Optimizer = sgd

Fig 4: Layers of existing ANN model

Source: Adapted from [3]

Proposed Work

About Dataset

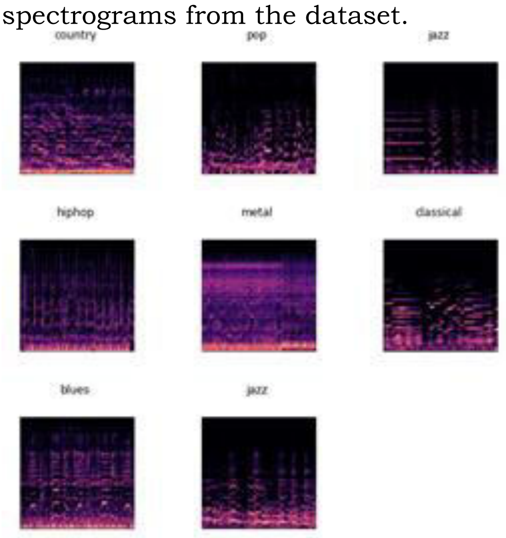
For training our model, we use the GTZAN dataset consisting of different types of data for the audio files. It has been widely used for music genre classification tasks. The data is split into 10 classes in this case genres of music, which are the following,

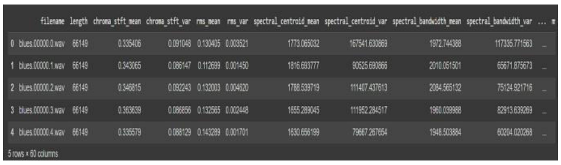
Blues
Jazz
Country
Metal
Pop
Rock
Reggae
Classical
Disco
Hip-Hop

Fig 5: Music Genres

The dataset consists of 1000(100 of each genre) audio files with 30 seconds of duration in .wav format. There are 2 .csv files that contain the audio files' various features in tabular form. One of the said .csv files consists for each song (exactly a 30 seconds sample) a variance and a mean calculated across various audio features. The second .csv file has the same structure, but the songs were split before into 3 seconds audio files (this way increasing 10 times the amount of data we fuel into our classification models). Spectrogram images are also present in the dataset in .jpg format with each genre consisting of 100 spectrogram images. These Mel spectrogram images were extracted from the audio files using audio pre-processing libraries. Below is a sample of 8 random Mel spectrogram images along with the genre they belong to.

Fig 6: A sample of a few Mel spectrograms from the dataset.





	filename	length	chroma_stft_mean	chroma_stft_var	rms_mean	rms_var	spectral_centroid_mean	spectral_centroid_var	spectral_bandwidth_mean	spectral_bandwidth_var	...
0	blues.00000.0.wav	66149	0.335406	0.091048	0.130405	0.003521	1773.065032	167541.630869	1972.744388	117335.771563	...
1	blues.00000.1.wav	66149	0.343065	0.086147	0.112699	0.001450	1816.693777	90525.690866	2010.051501	65671.875673	...
2	blues.00000.2.wav	66149	0.346815	0.092243	0.132003	0.004620	1788.539719	111407.437613	2084.565132	75124.921716	...
3	blues.00000.3.wav	66149	0.363639	0.086856	0.132565	0.002448	1655.289045	111952.284517	1960.039988	82913.639269	...
4	blues.00000.4.wav	66149	0.335579	0.088129	0.143289	0.001701	1630.656199	79667.267654	1948.503884	60204.020268	...

5 rows × 60 columns

Fig 7: A columns from the 3 second .csv file

Here, we will be using the 3 seconds .csv files containing 9990 rows and 60 features for the ANN classification task and all 1000 images for the CNN task.

For the tabular data of .csv format that contains features for each audio wave for a 3 second duration, we use a custom-made ANN after pre-processing the data using different methods. The ANN consists of 5 dense hidden layers all with a drop-out probability of and a different number of neurons. First, we apply the STFT or standard Fourier transform function on all the features in the dataset. It is used to determine the sinusoidal frequency and phase content of various different local sections of a signal over time as it changes. We then perform standard z score scaling to standardize the data and use librosa library to visualise the spectrograms of the audio files.

We split the data into training and testing set with a ratio of 0.2 such that the training set consists of 7992 records and the testing set contains 1998 records.

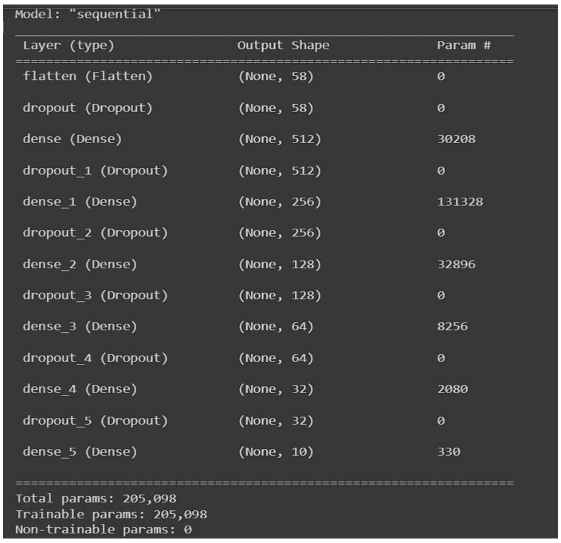
Model Building

Below, we propose two different models and approaches trained on the same data to perform this task. The models used are,

- ANN
- CNN (Transfer learning using EfficientNetB3)

First, we define the different algorithms that wereused for this paper

ANN: Artificial Neural Networks or ANN are fully connected or dense neural networks with multiple layers with an architecture shown in Figure 7. They are made up of an input layer, several hidden layers, and an output layer. Each neuron in a layer is connected to every other neuron in the next layer. This fully connected nature gives it the name dense. By increasing the number of hidden layers, we can make the network deeper.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 flatten (Flatten)           (None, 58)                0

 dropout (Dropout)           (None, 58)                0

 dense (Dense)               (None, 512)               30208

 dropout_1 (Dropout)         (None, 512)               0

 dense_1 (Dense)             (None, 256)               131328

 dropout_2 (Dropout)         (None, 256)               0

 dense_2 (Dense)             (None, 128)               32896

 dropout_3 (Dropout)         (None, 128)               0

 dense_3 (Dense)             (None, 64)                8256

 dropout_4 (Dropout)         (None, 64)                0

 dense_4 (Dense)             (None, 32)                2080

 dropout_5 (Dropout)         (None, 32)                0

 dense_5 (Dense)             (None, 10)                330

=================================================================
Total params: 205,098
Trainable params: 205,098
Non-trainable params: 0
```

Fig 8: Custom made ANN layers

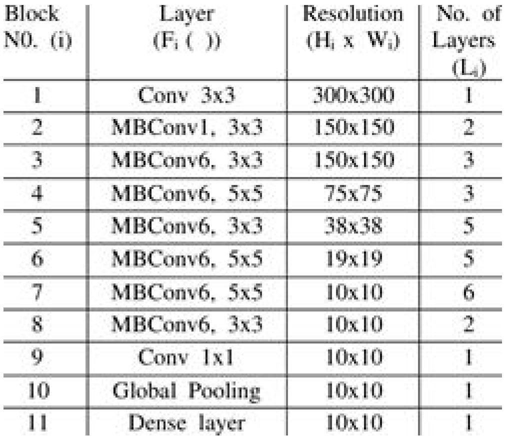
The flatten layer flattens the input with the shape. All layers have *ReLU activation function* and the output layer has the *Softmax activation function* which is useful in the case of multiclass classification problems. For the model that we developed, we have used *Adam* optimizer with a *learning rate of 0.001* and loss function as *sparse categorical cross entropy*.

In the implementation, we fed the model a batch size of 256 records at a time and ran the model for 600 epochs.

CNN: A Convolutional Neural Network also called a CNN is a deep learning algorithm that takes an image as input and assigns importance (learnable weights and biases) to different aspects/objects in the image and distinguishes one from the other by making feature maps. They consist of convolution layers that do the said task and pooling layers that reduce the dimensions of these produced feature maps. In the end, these feature maps are flattened and are passed to fully connected layers that learn from the passed feature maps and at last an output layer that consists of neurons equal to the number of classes. Compared to other deep learning algorithms that involve classification, CNN demands very less computation for pre-processing.

Transfer Learning: Transfer learning is a popular machine learning method in which a model that was originally trained for one task is used as a base for building a model on a different task. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems.

EfficientNet: EfficientNet is a CNN architecture and scaling method that uses a compound coefficient to uniformly scale all depth/width/resolution dimensions. The architecture of EfficientNetB3 is as follows:



Block N0. (i)	Layer (F_i ( ))	Resolution (H_i x W_i)	No. of Layers (L_i)
1	Conv 3x3	300x300	1
2	MBConv1, 3x3	150x150	2
3	MBConv6, 3x3	150x150	3
4	MBConv6, 5x5	75x75	3
5	MBConv6, 3x3	38x38	5
6	MBConv6, 5x5	19x19	5
7	MBConv6, 5x5	10x10	6
8	MBConv6, 3x3	10x10	2
9	Conv 1x1	10x10	1
10	Global Pooling	10x10	1
11	Dense layer	10x10	1

Fig 9: EfficientNetB3 layers

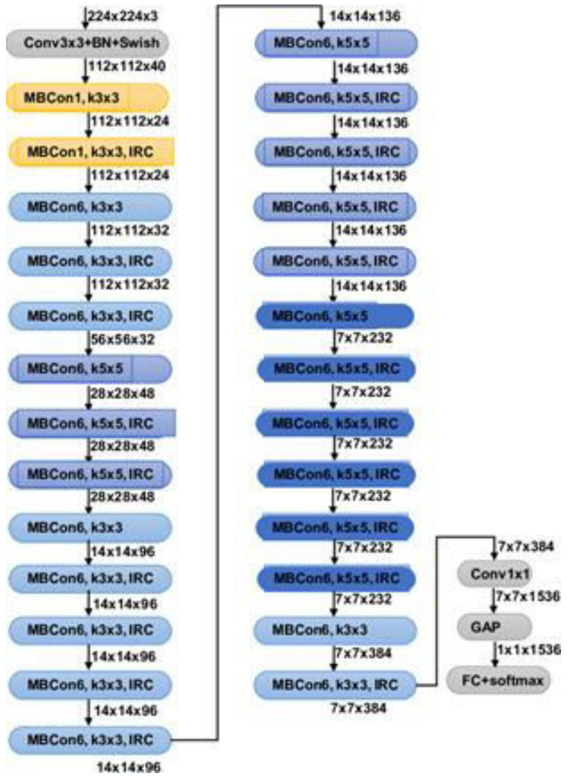


Fig 10: Architecture of EfficientNetB3 model.

Source: Adapted from [5]

In the case of classification based on Mel spectrogram images, we have 1000 images each class having 100 images. We split the data into training and testing with a ratio of 0.2 thus getting 800 images for training and 200 images for testing the model. Each image's dimension is set to (224,224,3) to fit into the model we are going to be using. We propose a model with the base model as EfficientNetB3 with initial weights of the ImageNetdataset.

We have also added a global average pooling layer and a dense output layer with 10 neurons having SoftMax activation function and some additional call backs along with the pre-trained model while training the model. The call-backs applied are early stopping and reduce learning rate. We used *Adam* optimizer with a learning rate of 0.001. We, then, fed the training images in batches of 8 at a time and ran it for 30 epochs.

Results and Discussion

Choosing our evaluation metrics is an important phase when it comes to predictions, especially classification problems.

Accuracy:

The accuracy denotes the sum total of correctly identified instances by the model divided by all the instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig 11: Formula for calculating accuracy

Spare categorical entropy loss function:

When true labels are one-hot encoded, categorical cross-entropy is applied. For instance, in the 3-class classification issue, the true values [1,0,0], [0,1,0], and [0,0,1] are available. Truth labels in sparse categorical cross-entropy are integer encoded, for instance, [1], [2], and [3] for a 3-class problem.

$$\text{Loss} = -\sum_{i=1}^{\substack{\text{output}\\\text{size}}} y_i \cdot \log \hat{y}_i$$

Fig 12: Formula for categorical cross entropy

Results for ANN task:

The training accuracy and the best validation accuracy came up to be 95.25%, 95.39% after training and evaluation.

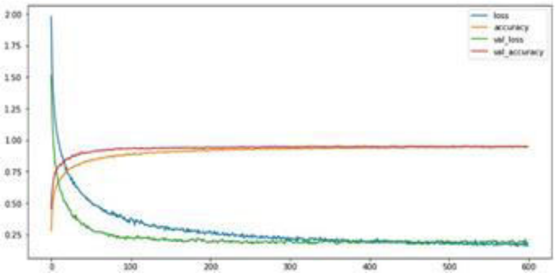


Fig 13: Graph of accuracies and losses versus number of epochs for the ANN.

Results for CNN task:

We got the highest validation accuracy of 84% and 99% training accuracy for the CNN task.

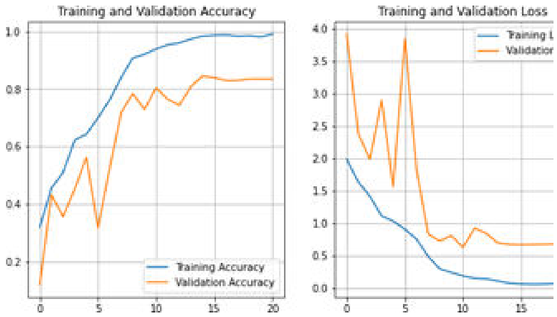


Fig 14: Graphs of accuracies and losses versus number of epochs for CNN.

Conclusion and Future Work

Deep learning techniques are extremely useful for classification tasks like music genre classification in which music is classified into different genres with respect to its features. The dataset used in this paper, GTZAN dataset, provides us with a rich variety of data in the form of audio files which are then pre-processed using various techniques like STFT which is a general-purpose tool used in audio processing. Using various libraries like librosa, we got tabular data and spectrogram images which are then fed into ANN and CNN respectively for classification. The models we used in this paper are simple but still proved to be very efficient when it comes to classifying music into different genres. The highest accuracy was brought out by our custom-made ANNmodel (95.35%) which is a very effective and widely used model whereas our CNN model broughtvalidation accuracy of 84%.

Future work can be done by using a larger dataset consisting of more records of audio files. In this paper, to increase the number of records, we split audio files of 30 seconds into 3 seconds. Dataset with a greater number of records/audio files can be used in the future for better performance and accuracy. Also, different combinations of different deep learning models (hybrid model) can be used to improve the complexity and accuracy of the model.

References

[1] Rajeeva Shreedhara Bhat, Rohit B. R., Mamatha K. R. “Music Genre Classification”, SSRG International Journal of Communication and Media Science (SSRG-IJCMS) – Volume 7 Issue 1 – Jan - April 2020

[2] Derek A. Huang huangda@stanford.edu Arianna A. Serafini aserafini@stanford.edu Eli J. Pugh epugh@stanford.edu

[3] Dr Jatinder Manhas. Rachit Kumar Gupta “Improved classification of cancerous histopathology images using color channel separation and deep learning”. Journal of Multimedia Information System VOL. 7, NO. 1, March 2020 (pp. 221-228): ISSN 2383-7632

[4] https://www.researchgate.net/figure/Illustration-of-the-EffecientNet-B3-architecture-10-3646- million-weights-IRC-means_fig4_348470984

[5] Image by Aqegg, https://en.wikipedia.org/wiki/Spectrogram

[6] Congyue Chen, Xin Steven “Combined Transfer and Active Learning for High Accuracy Music Genre Classification Method”, 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)

[7] Davis Moswedi, Ritesh Ajoodha “Music Genre Classification using Fourier Transform and Support Vector Machines”, 2022 International Conference on Engineering and Emerging Technologies (ICEET)

[8] S. J. Pan and Q. Yang, "A Survey on Transfer Learning", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[9] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, Qing He, “A Comprehensive Survey on Transfer Learning”arXiv:1911.02685[CS.LG]

[10] Joe Lemley; Shabab Bazrafkan; Peter Corcoran, “Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision”,IEEE Consumer Electronics Magazine(Volume: 6, Issue: 2, April 2017)