DOI: http://dx.doi.org/10.26483/ijarcs.v12i5.6771

Volume 12, No. 5, September-October 2021

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

MUSIC GENRE CLASSIFICATION USING NEURAL NETWORKS

Meet Raval
Dhirubhai Ambani Institute of Information and
Communication Technology,
Gujarat, India

Parv Dave
G. H. Patel College of Engineering and Technology,
Gujarat, India

Raj Dattani
G. H. Patel College of Engineering and Technology,
Gujarat, India

Abstract: In recent years, the complexity of making music has lessened, resulting in many individuals making music and submitting it to streaming media. Because of the huge music streaming media, people are spending a lot of time seeking for certain songs. As a result, the capacity to swiftly categorise music genres has become increasingly important. As machine learning and deep learning technologies progress, convolutional neural networks (CNN) are being employed in several fields, and several CNN-based versions have emerged one after the other. Traditional music genre classification necessitates professional abilities to manually extract features from time series data. We developed a music genre categorization model using CNN's audio advantages and features to save users time while searching for different types of music. During the pre-processing, Librosa is used to convert the original audio files into Mel spectrums.

The Mel spectrum is transformed and supplied into the suggested CNN model for training. On the GTZAN dataset, the 10 classifiers' decisions are subjected to a majority vote, with an average accuracy of 84 percent. Music genre categorization using neural networks (NNs) has seen some modest success in recent years. The success of song libraries, machine learning techniques, input formats, and the types of NNs utilised has all been mixed. This article looks at some of the machine learning approaches utilised in this sector. It also involves research on musical genre classification. Images of spectrograms produced from time slices of songs are fed into a neural network (NN) to classify the songs into different musical genres.

Keywords: Feature extraction, machine learning concepts, concurrent neural network, music genre classification, recurrent neural network, Genre recognition, ConvNet, Labelled Mel Spectrogram, Deep learning

I. INTRODUCTION

Since the Internet's birth, many people have uploaded music previously recorded on vinyl records or CDs to streaming media on the Internet. People look for popular music through music streaming services, which have become increasingly popular in recent years, and the enormous internet music library makes finding certain genres or songs difficult. As a result, a tool that can recognise and classify music is essential for novices and specific artists. Because most music in today's music streaming media just has a title and a creator, and most of it isn't tagged. This makes it harder to find hidden tags in songs and categorise them into genres. Machine learning has gained a lot of popularity in recent years. Depending on the type of application and the data set available, different types of machine learning algorithms are better suited for different applications.

There are four fundamental types of learning algorithms in machine learning. The four main forms of learning algorithms are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Unsupervised learning aims to extract important features from an unlabelled data set without a specific goal in mind, whereas supervised learning creates a mathematical model using a completely labelled data set. Semi-supervised learning, on the other hand, employs both labelled and unlabelled data in its data gathering. To adopt a new approach, reinforcement learning employs a feedback mechanism. Reinforcement learning is a type of learning in

which correct actions or predictions are rewarded. Reinforcement learning, for example, is frequently employed in games where the goal is to reduce risk while maximising reward.Genre is a nebulous notion, yet it is a defining feature of music. Automated genre categorization methods now in use generate a set of attributes from audio and create a classifier on top of them.

These properties are often calculated over a long period of time in such models. This paper presents a residual neural network-based model for genre categorization that was trained on short footage of only 3 seconds. Furthermore, standard genre categorization algorithms will assign an audio sample to a certain genre.

On the other hand, many genres are well-known for sharing characteristics. Given the ambiguity of the genre, the model proposed in this work may assign a music clip to one of three genre labels, each with a probability. Downloading and purchasing music from online music stores has become a common occurrence in the lives of a huge number of individuals across the world. Users frequently express their tastes in terms of genre, such as hip hop, pop, or disco. Most of the music currently accessible, however, are not automatically assigned to a genre. Automatic genre categorization is critical for music organisation, search, retrieval, and recommendation due to the large quantity of current collections. Due to the selection and extraction of relevant audio characteristics, music categorization is regarded as a difficult process. While unlabelled data is freely available, there are few music recordings with proper genre tags. The two essential processes in music genre categorization are feature extraction and classification. Various characteristics are retrieved from the waveform in the first stage. The characteristics retrieved from the training data are used to build a classifier in the second step. There have been a variety of ways to categorising music into distinct genres. When it comes to classifying songs in a music collection, a lot of human labour is required. People all around the world nowadays like and prefer to listen to songs of a specific genre. This function is available on many online music commercial sites; however, it is still done manually. The time-consuming procedure of first listening to a song and then selecting which genre it belongs to is no longer appropriate in the twenty-first century. To address this, we have attempted to suggest a new system for categorising music into genres.

In this work, we show that in a realistic setting, the classifier's predictions are consistent with a broader knowledge of genre. Music Genre Recognition is a well-studied subtask of Music Information Retrieval that has attracted manyresearchers from across the world.Identifying genre, on the other hand, is a challenging task in and of itself because there is no widely accepted definition of genre and significant overlap between different genre groupings. Developing an automated system for genre identification has been difficult, even though individuals can quickly recognise this abstract attribute with a music clip.The bulk of existing MGR techniques concentrate on collecting unique feature descriptors from music and then using these features to train a classifier. However, in the last two decades, deep learning, a powerful machine learning variant, has emerged, reducing, if not eliminating, the requirement for human feature extraction. Deep learning has become a powerful tool for developing robust end-to-end systems in the domains of image, video, audio, and speech analysis.

Music genre classification is extremely important today, due to the rapid growth of music tracks both online and offline. To have greater access to them, we need to correctly index them.Genre categorization is a method of grouping comparable sorts of data into a single identity (based on rhythm, instrumentation, or harmonic content) and assigning that identity to them. As the name suggests the music business is growing all around the world. Every day, fresh songs are composed. Since it was classified, every day, listening to such tunes would become a tiresome chore, were the technology may be used to cure music and make it more enjoyable. By utilising its rhythms, categorization becomes easier or more efficient.

Poetic composition and beats A song can be expressed in several ways. A signal in the form of an audio signal. This audio stream has a variety of characteristics such as root-mean-square, frequency, and spectral roll-off spectral centroid, (RMS) level, bandwidth, zero-crossing rate etc. The computer reads audio files in many formats, such as wav or mp3. This sound signal various music streaming applications, such as Spotify,Genre classification is used by Wynk, Apple Music, and others to recommend musicto its user's music. In the field of music genre, a lot of study has been doneclassification.These studies may be divided into three categories. There are two groups based on the dataset type. There are two well-known. There are two datasets available: FMA and GTZAN. The problems in this study are solved using a machine learning technique. Since the first convolution neural network was introduced, it has accelerated the field of deep learning in areas such as image classification and segmentation, object identification, and recognition. CNN is a type of neural network with a topology like that of a gird. This grid can be linear, such as for time series data, or 2D, such as for a picture.CNN decreases processing needs by employing a system comparable to a multilayer perceptron. For comparison analysis, support vector machines, artificial neural networks, multilayer perceptron’s, and decision trees are employed in addition to CNN.

A music genre classifier is a piece of software that determines the genre of music in audio format. These devices are utilised for things like automatically categorising music for services like Spotify and Billboard, as well as identifying acceptable background music for events. Now, genre classification is done manually by individuals using their own knowledge of music.

II. LITERATURE SURVEY

Tzanetakis and Cook presented music genre categorization as a pattern recognition problem in 2002. To evaluate categorization using acoustic features obtained from sound, the researchers utilised a dataset of 1000 music pieces divided into ten musical genres. The MARSYAS framework was also made available to the music information retrieval research community, and the authors published a comprehensive set of musical content characteristics. Since then, the MARSYAS framework has been frequently used to extract acoustic properties. Shen extracted different acoustic properties using the MARSYAS framework (timbre, rhythm, and pitch).

In a pre-processing step, the authors used Principal Component Analysis to decrease the dimensionality of the gathered feature vectors. After that, the features were combined to create a 25-dimensional feature vector. This feature vector was put into a three-layer neural network to accomplish nonlinear dimensionality reduction. The neural network output layer contains one unit for each dataset class.

The authors also incorporated human musical perception. The authors claim that their method is effective at recovering music and that it is novel in that it incorporates additional information (human perception) into a music retrieval test for the first time. In the idea of utilising complementary information in music genre classification, Song and Zhang proposed an information fusion framework for distance-based music genre classification algorithms. The findings of this fusion were better than those of a single feature set.Stochastic Gradient Descent was utilised by Sigtia and Dixon to train a neural network using rectified linear units.They used the activations of the hidden layers of the neural networks as features and trained a Random Forest classifier on top of them to forecast the classifications. The authors' hypotheses were tested using two datasets: GTZAN and ISMIR 2004. In the last 5-10 years, however, convolutional neural networks have proven to be incredibly accurate music genre classifiers, with excellent results reflecting both the complexity provided by having multiple layers and the ability of convolutional layers to effectively identify patterns within images. Our research revealed that when considering the whole 30s track duration, current state-of-the-art algorithms perform with an accuracy of around 91 percent, well above human skills for genre identification.

Gwardys and Grzywczak used a CNN trained on the Large-Scale Visual Challenge as a feature extractor (ILSVRC). They didn't have enough data in their scenario to train a classifier, but they did have enough data in a separate area of interest, where the data may be in a different feature space or follow a different data distribution.Pan and Yang showed that excellent knowledge transfer, also known as transfer learning, may significantly improve the learning algorithm's performance and reduce the requirement for expensive data-labelling procedures. For each image, a 4096-dimensional vector was created using the CNN, and three SVMs were trained, one for each image. Many of the papers that employed CNNs compared their models against other machine learning techniques including k-NN, Gaussian mixtures, and SVMs, and found that CNNs beatthem all.Our research revealed that when considering the whole 30s track duration, current state-of-the-art algorithms perform with an accuracy of around 91 percent, well above human skills for genre identification. Many of the papers that employed CNNs compared their models against other machine learning techniques including k-NN, Gaussian mixtures, and SVMs, and found that CNNs beat them all.We found some representation learning research in other applications, such as chord recognition and music starting detection, that is still within the scope of content-based music informatics.

According to the authors, utilising CNN for representation learning is a viable alternative to categorising short temporal characteristics and then smoothing the results into a musically plausible chord path via post-filtering. The results of their experiments back up their initial theory. Schluter and Bock looked studied Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) in the context of musical onset identification, which is the process of detecting when musically important events in audio data begin. They show that CNN outperforms RNN but at a higher computational cost due to the lack of human pre-processing. Using CNN for representation learning, according to the authors, is a feasible alternative to categorising short time features and then smoothing the findings into a musically credible chord route via post-filtering.

III. DATASET

The GITZAN dataset, which comprises 1000 music files, will be used. There are 10 different genres in this dataset, all of which have a similar distribution. Blues, classical, country, disco, hip-hop, jazz, reggae, rock, metal, and pop are among the genres included in the dataset. Each piece of music lasts 30 seconds.

Table1. Some Performance Results for GTZAN

Genre	Count
Blues	100
Classical	100
Country	100
Disco	100
Hip-hop	100
Jazz	100
Metal	100
Pop Rock	100
Reggae	100
TOTAL	1000

Music genre classification is a difficult process. However, technological advancements have mechanised this process, and millions of music programmes now categorise music into different genres in a matter of seconds. Above table summarises some of the categorization results for this dataset. The data collection contains 1000 audio recordings, each lasting 30 seconds. There are 100 tracks in each of the 10 genres (Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock). All the tracks are in.wav format and are 22050Hz Mono 16-bit audio files.

Table2. Performance Results for GTZAN

Reference	Accuracy
Tzanetakis	61%
Hozapfel	74%
Benetos	75%
Lidy	76.80%
Bregstra	83%

A few stages must be completed before the model may be built:

1. For computational efficiency, the mel spectrogram values should be scaled to be between 0 and 1.
2. The data presently consists of 1000 rows of 128 × 660 mel spectrograms. To show that there is just one colour channel, we need to rearrange this into 1000 rows of 128 x 660 x 1. We'd need this additional dimension to be 3 if our image contained three colour channels, RGB.
3. To be fed into a neural network, target values must be one-hot encoded.

The following is how the data was pre-processed:

1. A database was constructed for the whole collection and saved as a.csv file.
2. The libROSA library in Python is used to perform Feature Vector Extraction.
libROSA is a Python library for music and audio analysis that offers the foundation for creating music information retrieval systems.
3. Each audio file is extracted, and its feature vector is calculated. MFCC (Mel-Frequency Cepstral Coefficients) is the name of the extracted feature vector. By storing the approximate form of the log-power spectrum on the Melfrequency scale, the MFCCs encode the timbral characteristics of the music signal. Each audio track is represented by a Zero Crossings graph. The number of times the signal passes the zero level is depicted in this graph.
4. The music signal is subjected to Fourier Transforms. As a result, a Frequency Spectrum is obtained. Mel Frequency Spectrum is created by applying Mel Scale Filtering on the frequency spectrum.This Mel Frequency Spectrum is given a log () function before being converted into Cepstral Coefficients using discrete cosine transformations.

IV. METHODOLOGY

Feature Extraction
Audio characteristics serve as a quantitative method of conveying the most significant information in an audio clip. Feature extraction refers to the process of extracting important features included within raw data. This method transforms an audio input into a series of feature vectors. Feature extraction removes superfluous information from an audio stream and produces a more compact representation. According to the standpoint of music comprehension, audio characteristics may be separated into two levels: top-level and low-level.

The top-level labels give information on how listeners interpret and comprehend music using various genres, moods, instruments, and so on. Some significant feature extraction methods utilised in this investigation have been discussed in this section of the publication. Feature extraction approaches in music signal processing may be categorised in various ways. One of these is digital signal processing in the time and frequency domains. Statistical descriptors such as mean, median, and standard deviation are another effective method for feature extraction. All the techniques discussed below divide a raw music signal into N number of windows, and they are all run N times.

Based on their time scale, low-level audio characteristics may also be divided into short-term and long-term features. Audio characteristics include temporal domain features, frequency domain features, modulation frequency domain features, and so on. Linear Prediction Analysis, Spectral Centroid, Spectral Flux, Zero Crossing Rate Linear Predictive Cepstral Coefficients, Fast Fourier Transform, Mel scale Cepstral analysis, Spectral Roll-Off, low order statistics, and Delta coefficients are some feature extraction techniques. The MFCCs are the most widely utilised in voice recognition.

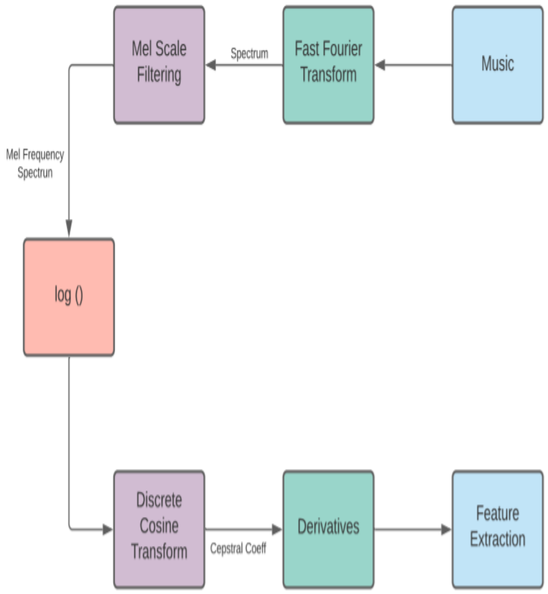

Fig1. Extraction of Feature Vector

MFCC are a widely used set of features in pattern recognition. MFCC was initially designed for automated speech recognition systems but has recently been utilised successfully in a variety of musical information retrieval applications. Because MFCC considers human perception sensitivity to frequencies, they are ideal for speech/speaker recognition. MFCCs are compact, short-time descriptors of the spectral envelope audio feature set that are typically computed for audio segments ranging from 10 to 100ms. The MFCC is a sound power spectrum with a short-term power spectrum. It is based on a nonlinear Mel frequency scale and a linear cosine transform of a log power spectrum. The magnitude coefficient of each short-term Fourier transform is multiplied by the appropriate filter gain, and the results are summed. To get MFCC, the discrete cosine transform is performed to the log of the Mel spectral coefficients.

Then, from audio files, we must extract significant characteristics. We will categorise our audio samples using five characteristics, namely Mel-Frequency. Cepstral Coefficients, Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, and Spectral Roll-off are all examples of spectral parameters. After that, all the characteristics are added to a.csv file so that classification methods may be utilised

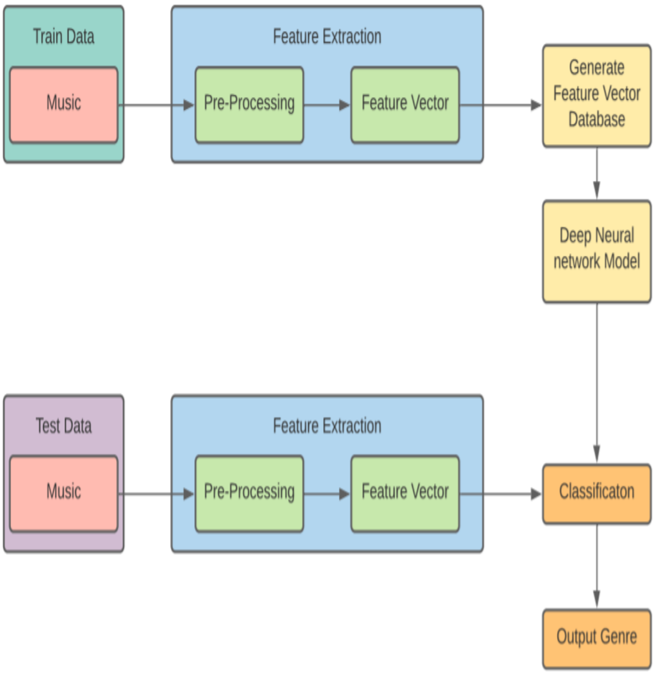

Fig2. Detailed Methodology

Methodology in detail:

1. There are two components to the dataset: training data and test data.

2. A feature vector is retrieved for each track in the train dataset after it has been pre-processed. The retrieved feature vectors are used to create a Feature Vector Database.

3. The acquired feature vector database is used to train the Neural Network model.

4. Each track in the test dataset is pre-processed and a feature vector is extracted.

5. To conduct classification on test data, the trained Neural Network model uses the feature vector produced at the conclusion of step 4 as input.

6. Finally, the output is the music track's genre.

Cepstral coefficients are linear scale. But human can hear the frequencies below 1 KHz as linear scale and the frequencies above as logarithmic scale. Because MFCC is one of the commonly used features in speech and speaker recognition systems. Frame Blocking, Windowing, Fast Fourier Transform, Mel Frequency Wrapping, and Spectrum are the steps of MFCC. Another parameter in the MFCC is the number of coefficients N. N is 13 in this investigation.

Table3. Music Genre Classification by Using Machine Learning

Feature	Statistical Functions	# of Features
Spectral Centroid	Mean, Median, Standard Deviation	3
Spectral Contrast		3
Spectral Roll off		3
Spectral Bandwidth		3
Zero Crossing Rate		3
MFCC(13 coeff)		39
MFCC Derivation		39
TOTAL		93

VI. Result and Discussion

The dataset contains 1000 audio recordings, each lasting 30 seconds. All the tracks are in.wav format. Among the genres featured are blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each category has 100 songs. We evaluated our models using this ten-class classification scheme.

We maintained a constant musical genre distribution in each fold, with 80% songs of each genre in the train split and 20% songs of each genre in the validation split.The discoveries lead to several fascinating conclusions. First and foremost, we discover that when paired with music transfer learning features, the multilayer perceptron model gives the best results.

This is understandable considering that the initial system was trained on the huge Million Song Dataset, which includes rich label sets for mood, age, instrumentation, and, most importantly, genre. Our experimental setting was also fine-tuned further, yielding the greatest results.This section examines the findings of all studies and compares them to the baselines of prior research. The CRNN model used in this study was trained using audio samples of various durations, split kinds, and feature levels.

More research on automated music genre categorization, as well as related research into other aspects of musical similarity, will be available in the future. The benefits of a large-scale effort to build high-quality ground truth would well outweigh the time and effort required.

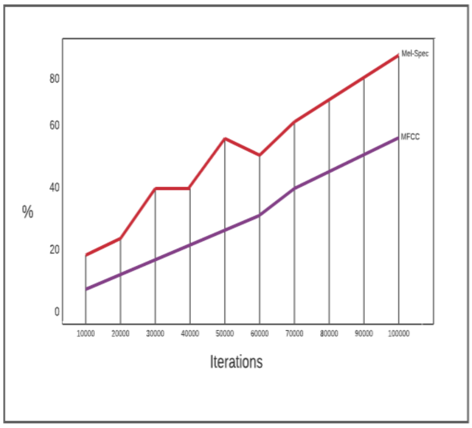

Fig3. Frequency %

VII. Conclusion

This study demonstrates an automated music genre categorization system based on Convolution Neural Networks. Mel Spectrum and MLCC are used to compute the feature vectors. The librosa package, which is written in Python, aids in the extraction of features and, as a result, in supplying suitable parameters for network training. As a result, this technique appears promise for categorising a large collection of music into the appropriate genre.After experimenting with various datasets, pre-processing methods, neural network architectures, and other parameters, we discovered that a convolutional neural network employing mel-spectrograms of three-second audio samples were the best combination.

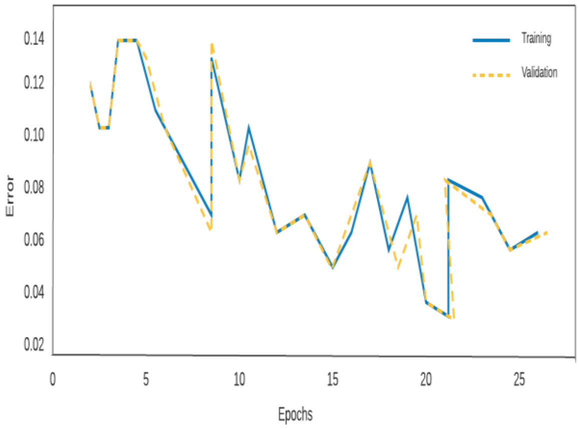

Fig4. Scoring History

A larger dataset decreases overfitting but has minimal effect on validationaccuracy. Although it fell short of the state-of-the-art accuracy for music genre categorization, it outperformed prior convolutional neural network attempts to address this problem. We also observed that categorization accuracy varied greatly by genre, which might have hampered overall performance.The objective of this study is to categorise and propose music using aural features gathered by digital signal processing methods and convolutional neural networks. The study was divided into two parts: studying how features for recommendation are collected and developing a service that recommends music based on user requests. Initially, feature extraction was accomplished using digital signal processing methods, and then CNN was taught as an alternative feature extraction method. Using acoustic features of songs in categorization, the optimal classification technique and suggested results are then identified.We will concentrate on this issue since deep learning techniques necessitate high-performance computer equipment. This research article describes an application that employs Machine Learning techniques to do Music Genre Classification. The programme does classification using a Convolutional Neural Network model. A Mel Spectrum is assigned to each track in the GTZAN dataset. This is accomplished using Python's libROSA package. A piece of software has been developed to categorise a vast database of music into its suitable genres. This study might be expanded to cover larger data sets and music in various formats.Furthermore, the style represented by each genre will change through time. As a result, the objective for the future will be to stay up to speed on changes in genre styles and to adjust our programme to accommodate these updated styles. This work may possibly be developed to function as a music recommendation system depending on a person's mood.

The method will be further developed in the future to classify music based on mood. This will be useful in determining which kind of music can alleviate stress in a person while listening to them. This can be used in music therapy, where a certain piece of music can be played based on the person's stress level. This work will need to be expanded for such a system.Future study on algorithm and feature engineering adjustments will involve tweaks to the initialization of theweights and testing with alternative filters while converting the mp3 file to a spectrogram.

VIII. References

[1] G.Tzanetakis and P. Cook. “Musical genre classification of audio signals, Speech and Audio Processing”, IEEE Transactions, July 2002.

[2] Vishnupriya S, and K.Meenakshi, “Automatic Music Genre Classification using Convolution Neural Network”, IEEE Conference 2018.

[3] Matthew Creme, Charles Burlin, Raphael Lenain, “Music Genre Classification”, Stanford University, December 15, 2016.

[4] Eve Zheng, Melody Moh, Teng-Sheng Moh, “Music Genre Classification: A N-gram based Musicological Approach”, 7th International Advance Computing Conference, 672-677, 2017.

[5] Daniel Grzywczak, Grzegorz Gwardys, “Deep image features in music information retrieval”, 10th international conference, AMT 2014, Warsaw, Poland, August 11-14, 2014 proceedings, pp 187-199.

[6] Chandsheng Xu, Mc Maddage, Xi Shao, Fang Cao, and Qi Tan,” Musical genre classification using support vector machines”, IEEE Proceedings of International Conference of Acoustics, Speech, and Signal Processing, Vol. 5, pp. V-429-32, 2003.

[7] Muhammad Asim Ali, Zain Ahmed Siddqui, “Automatic Music Genres Classification using Machine Learning”, International Journal of Advanced Computer Science and Applications, Vol 8, No 8, 2017.

[8] HareeshBahuleyan, “Music Genre Classification using Machine Learning Techniques”, University of Waterloo, ON, Canada, 2018

[9] Sam Clark, Danny Park, Adrien Guerard, “Music Genre Classification using Machine Learning Techniques”, 2012.

[10] T. Feng, “Deep learning for music genre classification”, 2014.

[11] Ahmet Elbir, Hamza Osman Ilhan, GorkemSerbes, Nizamettin Aydin. "Short Time Fourier Transform based music Genre classification”, 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018

[12] S. Sigtia and S. Dixon, “Improved music feature learning with deep neural networks,” in International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2014, pp. 6959–6963.

[13] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” arXiv preprint arXiv:1511.07289, 2015.

[14] Zhao, M. K. Hryniewicki, Z. Nasrullah, and Z. Li, “LSCP: Locally selective combination in parallel outlier ensembles,” in SIAM International Conference on Data Mining (SDM). SIAM, 2019.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv preprint arXiv:1412.3555, 2014.

[16] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[17] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” Journal of Machine Learning Research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[18] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International Conference on Machine Learning, 2015, pp. 448–456.

[19] H.G. Kim, N. Moreau, T. Sikora, “MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval,” John Wiley & Sons, 2005.

[20] Daniel Grzywczak, Grzegorz Gwardys- Warsaw University of Technology, “Deep Image Features in Music Information Retrieval”, August 2014, International Conference on Active Media Technology, DOI:10.1007/978-3-319-09912-5_16.