2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC)

A Survey on Music Genre Classification Using Multimodal Information Processing and Retrieval

Shreya G. Abhyankar, Shashank S. Bharadwaj, G. Shobha Rani, Pruthvi G. Karigiri*,
Sahana Srikanth and Sanjeev Gurugopinath
Department of Electronics and Communication Engineering, PES University, Bengaluru 560085, India
*Department of Computer Science and Engineering, PES University, Bengaluru 560085, India
Emails: {shreyagabhyankar, shashanksbharadwaj, shobharanig, pes1ug19cs354}@pesu.pes.edu, {sahanas, sanjeevg}@pes.edu

***Abstract*—Musical descriptors play a very crucial role in automatic music genre classification. With the increase in digitized music, query and retrieval of musical data has always been subjective and controversial. More often, a musical clip is often spread across the genres, making the classification process very challenging. This paper presents a contemporary survey of the recent advances made in the classification of music genre using different machine learning and deep learning algorithms. Specifically, it looks into the recent developments in the area of genre classification based on audio based descriptors, image based descriptors, and other descriptors which include album cover art, brain signals, lyrics and musical instrument digital interface data.**

***Index Terms*—Genre classification, deep learning, machine learning, multi modal information retrieval, music signal processing.**

I. Introduction

Music exerts a powerful influence on human beings. It is diverse in nature, which encompasses emotions, instruments, dance styles, gestures, and is culture specific [1]. Music, particularly the rhythm and the melody, works emotionally on all living creatures. It is known to simulate specific cerebral circuits in homo sapiens. Studies show multidimensional benefits on human life which include social bonding, cognitive and emotional behavior, stress reduction, mood elevation, memory stimulus, pain management and also sleep pattern betterment. With the rapid development in multimedia content, there has been an enormous increase in the accessibility of digital data to the general user. In the present times, a plethora of digitized music information has now become readily available to the general public. This has also resulted in a rise in music streaming across the world. Such large-scaled, abundant and diverse information requires careful annotation, which in turn needs beyond consistent human intervention, which has led to the development of automated retrieval of music information. These techniques include retrieving, collecting, browsing, storing, and recommending large musical data [2].

Music annotations have lately emerged as one of the popular areas of research. Most of the earlier methods involved human annotators which would need expertise, and were also time-consuming. The information present in a musical piece has many facets, and each of them has a key role to play in the annotation process. Few of such facets include audio signal, facial expression, lyrics, through the movements, temporal and perceptual characteristics, context and gestures made by a dancer or musician during his performance [3]. In addition to music being multi-faceted, it also faces the challenge of being influenced by the socio-cultural aspects. It also depends upon the perceiver's mood, circumstances and his/her experience [4]. It becomes essential to automatically discriminate the musical piece based on their rich intellectual facets and improve its accessibility to the general users.

Automatic classification of music based on genre has evolved as the most efficient way to separate different types of music ranging from classical, jazz, rap to pop [5]. Such labels are helpful in classifying musical compositions into general categories which have common temporal, musical or regional characteristics. However, handling such large sets of music introduces various challenges for information retrieval and classification. Streaming services often possess catalogs of huge collection of music pieces, for whom genre classification is of primary importance [6]. Many scientists and engineers have tried to automate this process by analyzing the human perception of music. The various descriptors such as tempo, chords, melody, pitch, rhythm, timbre, spatial features etc. help formalize the process of music classification. This has given rise to developments in the area of music processing. This includes machine listening of music or deep neural extraction of structures and patterns in the musical piece.

This paper discusses an extensive survey on the different deep learning (DL) and machine learning (ML) techniques used for classification of music based on genre. It starts with the importance of genre based music classification and goes on to discuss the different intellectual descriptors present in the music. This survey is based on three different categories: (a) audio-based, (b) image based and (c) others, which include modalities different from audio and images. Specifically, in audio-based classification, this papers consider the physical and perceptual characteristics such as texture, timbre, pitch and rhythm. On the other hand, image-based classification is based on visual representation of music based on album cover art, spectrograms, heat maps, colors, histograms etc. Additionally, there's also discussion on some other information used in the genre classification, such as brain waves, textual content including song metadata which includes information such as artist name, critic reviews, etc.

This review paper is organized as follows: In Section II,

979-8-3503-9619-5/23/$31.00 ©2023 IEEE

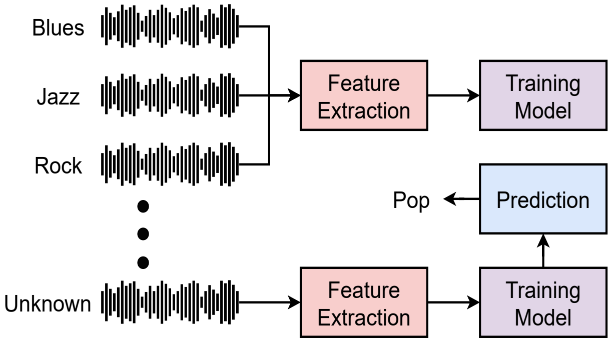

Fig. 1: Audio-based genre classification.

the paper gives a brief overview of the genre-based music classification. In Section II-A, there's discussion on the audio-based music genre classification based on timbre and texture, rhythmic and pitch-based features. The image-based genre classification is discussed in Section II-B. Some of the other features the researchers have looked into to classify the music genres are highlighted in Section II-C. It also discusses some of the publicly available data sets in Section III, and conclusions are provided in Section IV.

II. Genre Classification

Music genres are high-level descriptors or labels used by humans to organize and categorize the vast music collection, albums and artist. Music is often characterized based on certain characteristics such as style or type of rhythm, structure, harmonic content and instrumentation into categories such as Hip-hop, Jazz, Country, Rock and many more. This helps in easy organization of large musical databases available. Earlier, humans used to listen to a track and classify it into genre it belongs to, [7]. The automation of this process reduces human errors and time to a great extent. In this section there's consideration of the developments that have taken place in the recent past, describing the various modalities that were used to collect the musical information. These may include audio, image, lyrics, symbolic scores, album covers, text and so on. These modalities are obtained at different times through different transducers at various locations and/or belong to different media. This paper considers genre based music classification in three different categories, namely (i) Audio based features, (ii) Image based features, and (iii) Other modality based features.

A. Audio-Based Features

Audio features are based on physical and perceptual context. These features include timbre, melody, harmony, rhythm, pitch, tonality, dynamics, spatial location, semantic features, short-term and long-term features, and composition based features. Each of these descriptors capture different aspect of the audio signals, [8]. These descriptors are based on time-frequency representation of the audio signal and are widely used in the area of speech processing. The Fig. 1, represents an overview of the process involved in audio based classification. The different short and long term features are extracted from the audio file to represent the feature set. These features are given as an input to the classifier to train the model based on different genres which is then used to predict the genre of an unknown musical piece. This section looks into the developments in the audio based music genre classification. We categorize these developments into three major categories:

1) Timbre and texture-based features.
2) Rhythm-based features.
3) Pitch-based features.

1) Timbre and Texture-Based Features: The timbre-textural features are based on the Mel-frequency cepstral coefficients (MFCC) and the short time Fourier transform (STFT). The analysis are based on small segments of audio files, each processed independently. Many timbre-texture features such as time-domain zero crossings, roll-off of spectral centroid and spectral flux. Mel-cepstral coefficients are calculated for small sections. Each of these sections are processed by applying a suitable window function to extract the frequency content in the signal. Different window functions show specific behaviour in the frequency domain. The type of the window, size of the window and the overlap ratio greatly effect the classification accuracy. In [9], the authors perform a detailed analysis on the effect of six different window functions such as Rectangular, Parzen, Blackman, Kaiser, Hamming, Bartlett, with different sizes and overlap ratio using support vector machine (SVM) algorithm using three kernel functions namely, linear, polynomial, random forest (RF) and radial basis function (RBF) algorithms. Many long term features of the audio files are often captured considering the long segments of audio files such as mean, variance, average root mean square (RMS) energy. The authors in [10] have implemented genre classification using SVM and K-nearest neighbour (KNN) algorithms. The proposed method finds a region of interest by decomposing the music signal by employing the Empirical Mode Decomposition (EMD). Further, time-freqeuncy related statistics such as mel frequency coefficients, median frequency, excess kurtosis, spectral roll-off, zero crossing rate, standard deviation etc. are computed. The accuracy scores were based of different SVM kernels, namely linear, quadratic, cubic, Gaussian and KNN kernels such as fine and medium Kernel. In [11], a Mel spectrum and MFCC of the audio signals were used. The paper compares the learning accuracies of these feature vectors using a convolutional neural Network (CNN). In [12], different ML algorithms have been applied. This paper considers MFCCs of each audio as the feature vector. Naive Bayes, Decision Tree and KNN with k=5 used. In [13], paper presents a comparison on SVM, long short term memory (LSTM) and a proposed hybrid LSTM model based on SVM to classify different genres. However, the classification based on feature extraction relatively takes longer duration of audio clips. In addition to this, ambiguous genre have overlapping characteristics. To overcome this, a residual neural network was proposed in [14], where the training time was significantly reduced with short duration of 3 seconds audio clips. The model also assigns top-

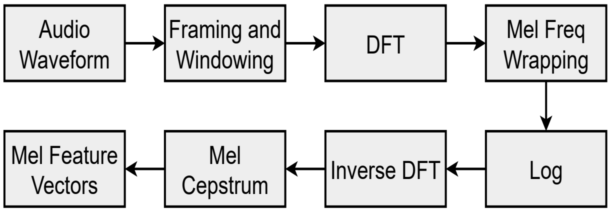

Fig. 2: Audio signal representation and preprocessing.

three different genre labels to a audio clip with highest mean probabilities based on soft voting. In [15], the authors extract the tempo gram as a feature to classify the audio clip using Gaussian mixture models. The shape of spectrum in every acoustic class is characterized by the first two moments. [16] explores various ML and DL techniques for the classification task. Ensemble models have been considered too. CNN model gives the leading metric in this study. In [17], the authors have investigated the fact that extracting features vectors from individual music instruments which may be genre specific can classify genre from a large database. The parameters from both original audio and separated tracks have improved the classification accuracy to a considerable extent. In Fig. 2 , the audio signal representation approach of converting a waveform to MFCC/Mel spectrograms is depicted.

2) Rhythm-Based Features: The rhythm in the music captures the occurrences of notes and silences in the audio file. It gives an estimate of the main beat and the sub-beats in the audio file and their relation and their relative strength. The histograms are widely used to capture the rhythmic content in the audio file. Some of these features include histogram peak amplitudes, their ration, period of these peaks, and overall histogram sum. The beat histograms are used to extract musical parameters such as tempo and statistical properties such as mean, variance or standard deviation. In [18], the authors have used the rhythmic content based on beat histograms to extract the feature vector and extreme learning machine (ELM) with bagging and KNN classifier were used to classify the genres. In [19], the amplitude, generic and tonal changes in the signal spectrum are captured to get the rhythmic information in the signal. The authors have analysed the feature set to identify the least and the most informative features through selection techniques and have used SVM for classification into five different genres. The authors in this paper have also analysed the base line feature set and have brought out the comparison against this feature set. In [20], bass lines were used to classify the music into different genres. Bass lines capture both harmonic and rhythmic aspects of the music. The bass line features include, ratio of the strongest peak power or the next strongest peak power to the total power, period of the highest or the next-highest peak, sum of powers of all the frames in the window, which are often used along with the timbral features to improve the classification accuracy. Principal component analysis was used to reduce the feature space dimensionality, along with linear component analysis (LDA) and classification of the given genre was done using Bayes decision rule. Lately, moment in-variants have been used to capture both rhythmic and tonal content. In [21], authors have proposed Hu moment based features extraction using sonogram and chromagram to classify the music audio into 5 different genres based on energy concentration in the spectrogram. The authors exploit the correlation between the coefficients of the Mel filter's between the adjacent frames and within the frames which are not a part of MFCC and brings out the comparison between the Hu-based feature set and generally used MFCC feature set using SVM.

3) Pitch-Based Features: Pitch is a primary auditory attribute in the music among timbre, texture, and rhythm. It forms a robust attribute for genre classification as it remains immune to frequency modulation. The accurate multiple pitch detection have been a challenging task. However, statistical information in the music can be extracted using pitch histograms. In [22], pitch histogram features such as the bin number of the strongest peak, amplitude of the strongest peak and interval in between the two strongest peaks of the folded histogram, and period of the maximum peak of the unfolded histograms are calculated to classify the genres into five classes using KNN. The authors in [23], have used a multi layer perceptron to classify the audio file into different genres. The feature vector extracted in this paper are based on energy distribution locally across various pitch frequencies by first decomposing the signal to extract the signal of interest by ignoring a very high and low frequency characteristics. In [24], the classification into music, noise and speech was carried out based on pitch. The feature set were carefully selected using forward selection procedure which include standard deviation within the classification window, the absolute difference between the subsequent samples in pitch, tonal distance and difference between the strongest and weakest pitch were used in a probabilistic model with softmax activation function. The music from different genres are often taken at different scales. In addition to this, environmental sounds make it difficult to distinguish genres based on the pitch. In [25], the author explores pitch and speech related algorithms to describe singing and understand gliding, portamento and vibrato. The "pitch class profiles", which is also referred as Chroma based features can also be used to classify music into different categories. Human auditory system perceives notes which are one octave apart to be similar. In [26], the authors have used chroma feature which is usually 12-dimensional vector on a VGG16 deep learning network to classify the genres in a music file.

B. Image-Based Genre Classification

Visual representation have always attracted humans through colors, graphs, charts, heat maps and many other ways. Human brain are equipped to easily comprehend and process patterns/trends in visual information. It is not straight forward to apply time domain features to the neural networks due to its high sampling rate, [27]. The music spectograms have proved to be alternative to this and captures both time and frequency content. The spectograms are visual representation of different frequencies in signal since it is a function of time.

The colormaps are used to identify the magnitude of these frequencies within the time windows. In [28], 10-second sound clips of You Tube videos were used to classify the genre using a CNN based VGG-16 neural classifier. The audio file are extracted from this which include instrument, vehicle, speech and animal sound along with the music. The 2D spectograms are extracted and are considered as images to be given as input to the neural networks. The musical spectograms can also be used to speed up the learning process. The method of transfer learning has been lately used in literature, where in the few layers are already trained to learn the core features are removed and new model-specific layers are introduced to focus on the fine tuning of the neural network. This method of transfer learning greatly reduces the training time and computational requirements. The authors in [29], have used a transfer learning approach to reduce the miss classification rate of broad genres such as Pop (describes popular music) and R&B from the other genres for western music. The transfer learning approach uses two different tasks: Target and source tasks. In source task, the pretrained models were used in this paper for classification of 1100 audio recordings into 11 genres in the Target task. In [30], four different transfer learning algorithm, Resnet34, Resnet50, Alexnet and VGG16 were used to classify into 10 genres. In [31], the attention based serial and parallel bidirectional neural networks were used with audio STFT spectrograms as input to the neural networks. Deep residual networks (ResNets) consists of bypassing two or more convolution layers, which are common feed-forward CNNs with residual learning. In order to capture the low level information of mel spectograms, the authors in [32] use bottom up broad cast neural networks with inception blocks having different kernel sizes, connected densely to transfer low level information to higher layers. These information may include texture, pitch, loudness, etc. This method greatly increases the decision making as low level descriptors are also available now for decision making.

In [33], results on spectograms obtained from CNN are compared with that possessed due to SVM classifiers and prepicked features. Additionally, fusion of learned and hand-crafted features are carried out, and corresponding results are discussed. Most of the CNNs fail to capture long-term temporal information.Therefore In [34], a parallel recurrent convolutional neural (PRCNN) which consists of a Bi-RNN and CNN for spatial feature extraction and classification has been used. In [35], the genres were classified based on color histograms. ML techniques, namely, lasso, ridge, SVM, adaboost, KNN, random forest and discriminant analysis were used to make classification.

The various color histogram are also used as digital images to classify the music genres. They capture the distribution of different colors in an image. A histogram presents the number of pixels of particular color present in an image. It just indicates the proportion of different colors present irrespective of its spatial location in an image. In [36], a similarity score of artists which is used to annotate them within a set of genre classes is developed. This technique, called joint equal contribution (JEC) is based on album art or promo photographs, makes use of texture and multiple forms of low-level color (color histograms - HSI, RGB). It is further established on standard data sets that JEC outperforms several state-of-the-art techniques. In Fig. 3, shows one such presentation of different genres. The horizontal axis represents the pixel values divided into different bins with each bin representing a small range of pixels and the vertical axis has the number of pixels in that range. The Fig. 3, represents the color histogram of four different genres. The authors in [6], multimodal representation based approach to classify the music genre using heatmaps, text and audio. In this paper, the CNN is used to generate heatmaps and the authors have demonstrated what features are highlighted in the heatmaps for different genres.

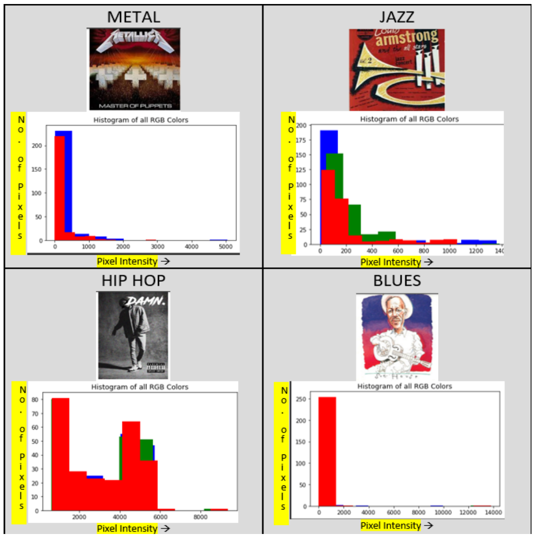

Fig. 3: Histogram representation of genres.

C. Other Techniques

There are many music descriptors such as text, album cover art, album, artist and brain waves etc, which may be considered for music genre classification, they often classify based on personal preferences. In [37], authors have classified music genres based on human preferences. The magnetoencephalogram and electroencephalogram brain signals were used for the music clips classification. The lyrics of the song often give a hierarchical structure representation, words to lines, lines to segments, segments to complete song. In [38], [39], Hierarchial attention network (HAN) and LSTM were used for genre classification based on lyrics. The prepicked features are used extensively for pattern description in songs. In [40], various handcrafted features from audio set such as robust local binary patterns (RLBP), statistical spectrum descriptors (SSD) and MFCC. Simplified chord sequences extracted from chords and lyrics were considered and detailed analysis on the performance was brought out through SVM, KNN, decision tree, random forest and representative learning algorithms such as CNN and LSTM. In [39], [41], [42], different models such

Ref.	Modality	Data Set	Method	Accuracy (%)
[10]	Audio	GTZAN-5 Genres	EMD, SVM-L	94
[13]	Audio	GTZAN	Hybrid SVM-LSTM	89
[16]	Audio	GTZAN	XGBoost CNN	79.3 86.2
[29]	Image	MSD	CNN	89.38
[30]	Image	GTZAN	Transfer Learning-Resnet34	79
[34]	Image	GTZAN, Extended Ballroom	Parallel Recurrent CNN	90.2 92.5
[39]	Lyrics	Private	HAN, BoW,LSTM, CNN Ensemble	66.9
[43]	Audio, Image, Text	MuMu	CNN,ResNets,VSM	93.6

TABLE I: Performance of leading methodologies across modalities.

as NGram, bag of words (BoW) and structural and statistical text features (SSTF) were used to classify the music based on Turkish lyrics. Music genres often share common boundaries. In [43], a multi-label genre classification was proposed based on audio, text, and images collected over 31000 albums with cover image, text reviews, and audio tracks. The album cover art is one of the visual style of experiencing a music. It gives a symbolic representation and an overview to the ingredients present in the musical album. For example, Pop music albums have solo artists or group in their front cover, and heavy metal albums usually have dark and non-regular fonts [44]. The music can also classified into different genres using symbolic representation, such as humdrum, music encoding initiative (MEI), MusicXML, musical instrument digital interface (MIDI), and so on. In [45], MIDI files were used and converted to MIDI images to be fed for music genre classification into a 3D convolutional denoising autoencoder, considering musical features such as pitch, bar length and volume.

One of the common techniques used in genre classification to reduce the computational complexity is music summarization. A section that possess its representative features is termed as a summary of that music data section. Conventionally, a corresponding algorithm first divides the music section into smaller frames to calculate feature vectors in each frame. Later, frames having similar features are clustered together and are given a label. Most frequency label-carrying longest section is identified as its corresponding summary, unlike identifying most frequently repeated section. Such a performance study has been reported in [46], where summaries from several music data sections are extracted and the performance is compared with the technique that uses single section summary extraction. This study shows that a better performance can be achieved by intelligent summary of music data. In particular, this is achieved by choosing the ratio and number of sections appropriately. Table. I represents the performance of the techniques that have leading accuracy across various modalities.

III. Data Sets

In this section, we list the publicly available dataset in each of the above categories.

1) GTZAN dataset: This dataset has audio tracks of 10 genres, each having duration 30 seconds. The audio files are in .wav format [47]
2) The million song dataset (MSD): It is a set of a million popular contemporary music. The dataset is available in the form of collection of audio features and meta data [48]
3) MTG-Jamendo dataset: This dataset is has more than 55,000 audio tracks with auto-tagging of genres, instruments and mood/theme tags. The audio tracks are in .mp3 format [49]
4) YouTube-8M dataset: This is a video based dataset and has more than million YouTube video. These videos are annotated with 4800 visual entities [50].
5) MuMu Dataset: This is a multimodal dataset, with genre annotations done over multiple labels that fuses information from a dataset based on reviews on Amazon and the million song dataset (MSD) [6].

IV. Conclusion

In this paper, we have presented a review on various learning based algorithms used to classify music into different genres based on audio, image and other modalities. We discussed the importance of genre classification and different descriptors used in the music that help in classifying music based on genres. Specifically, we mentioned different audio-based features and have looked into developments in audio genre classification, particularly in three categories, namely, timbre and texture, rhythm and pitch. Additionally, we presented image-based genre classification based on various visual representation of images. Lastly, we summarized other modalities such as text, album cover art, brain signals, symbolic representation to classify the genres. Our discussion also includes the various dataset publicly available in each of these areas.

References

[1] D. Hesmondhalgh, *Why music matters*. John Wiley & Sons, 2013.

[2] C. C. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic, "The need for music information retrieval with user-centered and multimodal strategies," in *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, 2011, pp. 1–6.

[3] S. Essid and G. Richard, "Fusion of multimodal information in music content analysis," in *Dagstuhl Follow-Ups*, vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[4] J. S. Downie, "Music information retrieval," *Annual review of information science and technology*, vol. 37, no. 1, pp. 295–340, 2003.

[5] C. McKay and I. Fujinaga, "Automatic genre classification using large high-level musical feature sets." in *ISMIR*, vol. 2004, 2004, pp. 525–530.

[6] S. Oramas, F. Barbieri, O. Nieto Caballero, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[7] S. Lippens, j.-p. Martens, T. De Mulder, and G. Tzanetakis, "A comparison of human and automatic musical genre classification," vol. 4, 05 2004.

[8] P. Manning, *Electronic and computer music*. Oxford University Press, 2013.

[9] A. Elbir, H. O. İlhan, G. Serbes, and N. Aydın, "Short time fourier transform based music genre classification," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*. IEEE, 2018, pp. 1–4.

[10] E. Chaudary, S. Aziz, M. U. Khan, and P. Gretschmann, "Music genre classification using support vector machine and empirical mode decomposition," in *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. IEEE, 2021, pp. 1–5.

[11] S. Vishnupriya and K. Meenakshi, "Automatic music genre classification using convolution neural network," in *2018 international conference on computer communication and informatics (ICCCI)*. IEEE, 2018, pp. 1–4.

[12] V. Prashanthi, S. Kanakala, V. Akila, and A. Harshavardhan, "Music genre categorization using machine learning algorithms," in *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*. IEEE, 2021, pp. 1–4.

[13] P. Fulzele, R. Singh, N. Kaushik, and K. Pandey, "A hybrid model for music genre classification using lstm and svm," in *2018 Eleventh International Conference on Contemporary Computing (IC3)*. IEEE, 2018, pp. 1–3.

[14] D. Bisharad and R. H. Laskar, "Music genre recognition using residual neural networks," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 2063–2068.

[15] R. Thiruvengatanadhan, "Music genre classification using gmm," *International Research Journal of Engineering and Technology (IRJET) e-ISSN*, pp. 2395–0056, 2018.

[16] V. Shah, A. Tandle, N. Sharma, and V. Sheth, "Genre based music classification using machine learning and convolutional neural networks," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021, pp. 1–8.

[17] A. Rosner and B. Kostek, "Automatic music genre classification based on musical instrument track separation," *Journal of Intelligent Information Systems*, vol. 50, no. 2, pp. 363–384, 2018.

[18] B. K. Baniya, D. Ghimire, and J. Lee, "Automatic music genre classification using timbral texture and rhythmic content features," in *2015 17th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2015, pp. 434–443.

[19] A. Lykartsis and A. Lerch, "Beat histogram features for rhythm-based musical genre classification using multiple novelty functions," 01 2015.

[20] T. Kitahara, Y. Tsuchihashi, and H. Katayose, "Music genre classification and similarity calculation using bass-line features," in *2008 Tenth IEEE International Symposium on Multimedia*, 2008, pp. 574–579.

[21] R. Lopes, S. Chapaneri, and D. Jayaswal, "Music features based on hu moments for genre classification," in *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*, 2017, pp. 22–27.

[22] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histograms in audio and symbolic music information retrieval," *Journal of New Music Research*, vol. 32, no. 2, pp. 143–152, 2003.

[23] R. Sarkar and S. K. Saha, "Music genre classification using emd and pitch based feature," in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 2015, pp. 1–6.

[24] A. B. Nielsen, L. K. Hansen, and U. Kjems, "Pitch based sound classification," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3. IEEE, 2006, pp. III–III.

[25] R. Ajoodha, R. Klein, and B. Rosman, "Single-labelled music genre classification using content-based features," in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015, pp. 66–71.

[26] L. Shi, C. Li, and L. Tian, "Music genre classification based on chroma features and deep learning," in *2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP)*. IEEE, 2019, pp. 81–86.

[27] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[28] H. Bahuleyan, "Music genre classification using machine learning techniques," *arXiv preprint arXiv:1804.01149*, 2018.

[29] B. Liang and M. Gu, "Music genre classification using transfer learning," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2020, pp. 392–393.

[30] J. Mehta, D. Gandhi, G. Thakur, and P. Kanani, "Music genre classification using transfer learning on log-based mel spectrogram," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2021, pp. 1101–1107.

[31] Y. Yu, S. Luo, S. Liu, H. Qiao, Y. Liu, and L. Feng, "Deep attention based music genre classification," *Neurocomputing*, vol. 372, pp. 84–91, 2020.

[32] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7313–7331, 2021.

[33] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, vol. 52, pp. 28–38, 2017.

[34] R. Yang, L. Feng, H. Wang, J. Yao, and S. Luo, "Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices," *IEEE Access*, vol. 8, pp. 19 629–19 637, 2020.

[35] N. Lee and R. Baraldi, "CSE 546 final paper predicting musical genre from album cover art."

[36] J. Libeks and D. Turnbull, "You can judge an artist by an album cover: Using images for music annotation," *IEEE MultiMedia*, vol. 18, no. 4, pp. 30–37, 2011.

[37] P. Ghaemmaghami and N. Sebe, "Brain and music: Music genre classification using brain signals," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 708–712.

[38] A. Tsaptsinos, "Lyrics-based music genre classification using a hierarchical attention network," *arXiv preprint arXiv:1707.04678*, 2017.

[39] C. L. Riyoichi Sawada Ueno and D. Furtado Silva, "On combining diverse models for lyrics-based music genre classification," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 2019, pp. 138–143.

[40] R. M. Pereira, Y. M. Costa, R. L. Aguiar, A. S. Britto, L. E. Oliveira, and C. N. Silla, "Representation learning vs. handcrafted features for music genre classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[41] Ö. Çoban and G. T. Özyer, "Music genre classification from turkish lyrics," in *2016 24th Signal Processing and Communication Application Conference (SIU)*. IEEE, 2016, pp. 101–104.

[42] Ö. Çoban and I. Karabey, "Music genre classification with word and document vectors," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2017, pp. 1–4.

[43] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text, and images using deep features," *arXiv preprint arXiv:1707.04916*, 2017.

[44] J. Q. Li, "Genre classification via album cover," 2020.

[45] L. Qiu, S. Li, and Y. Sung, "3d-dcdae: Unsupervised music latent representations learning method based on a deep 3d convolutional denoising autoencoder for music genre classification," *Mathematics*, vol. 9, no. 18, p. 2274, 2021.

[46] Y. KIKUCHI, N. AOKI, and Y. DOBASHI, "A study on automatic music genre classification based on the summarization of music data," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 705–708.

[47] B. L. Sturm, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[48] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere, "The million song dataset." 01 2011, pp. 591–596.

[49] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *ICML 2019*, 2019.

[50] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.