Educational Administration: Theory and Practice
2024,30(1), 4718 - 4724
ISSN:2148-2403
https://kuey.net/

Research Article

Enhancing Music Classification: Machine Learning Approaches to Song Type Recognition

Paromita Das[1*], Somsubhra Gupta[2], Biswarup Neogi[3]

[1]PhD scholar, MAKAUT, West Bengal, INDIA, Amity University Kolkata, West Bengal, INDIA. (rchparomita@gmail.com)
ORCID ID: 0009-0003-0189-6825
[2]Swami Vivekananda University, Barrackpore-700121, INDIA (gsomsubhra@gmail.com) ORCID ID: 0000-0003-2272-3915
[3]JIS College of Engineering, Kalyani -7412356, INDIA (biswarupneogi@gmail.com) ORCID ID: 0000-0002-0981-5383

Citation: Paromita Das· et al. (2024), Enhancing Music Classification: Machine Learning Approaches to Song Type Recognition, *Educational Administration: Theory and Practice*, *30*(9), 4718 - 4724
Doi: 10.53555/kuey.v30i1.8148

ARTICLE INFO **ABSTRACT**

In this paper, a new method for identifying Indian music under the machine intelligence framework is presented using digital signal processing (DSP) techniques. To extract useful information from audio signals, the suggested method makes use of aspects specific to Indian music, like pitch, tempo, and spectral characteristics. Several DSP methods are used to efficiently handle the audio data, such as spectrogram analysis and Fourier transformations. Furthermore, machine learning models are used for pattern recognition and classification tasks. Examples of these models are deep neural networks and support vector machines. Accurate identification of Indian music is made possible by the combination of DSP and machine intelligence, especially in the face of noise or stylistic changes in performance. The suggested method is effective in identifying Indian music, as shown by the experimental results, which also highlight its potential uses in content indexing, music recommendation systems, and cultural preservation initiatives.

Keywords: Digital signal processing, Machine intelligence, Machine Learning, Pattern recognition, Song type.

I. Introduction

A limited number of features were employed by the early audio classification algorithms; they included features related to intensity and zero crossing rate (Saunders, 1996; Scheirer & Slaney, 1997; Lambrou et al., 1998; Liu et al., 1998), while spectral features linked to timbre were soon added (Wold et al., 1996; Foote, 1997; Soltau et al., 1998). Early speech/music classification systems (Carey, Parris, & Lloyd-Thomas,1999; Dannenberg, Thorn, & Watson, 1997; Zhang, 1998) also incorporated pitch-related information, such as qualities of the fundamental frequency variation, but these features are typically not employed by systems that target polyphonic audio input. The number of features increased steadily over time, finally covering nearly every characteristic and adding even more instantaneous features (McKinney & Breebart, 2003; Li, 2000). The most popular features were still intensity and timbre; however, additional feature dimensions were added to the list of features, even though the classification results using these features appear to be fairly good. These extra features include basic tonal features like pitch histogram features (Tzanetakis & Cook, 2002; Burred & Lerch, 2004; Dixon, Pampalk, & Widmer, 2003), stereo panning features (Tzanetakis, Ermolinskyi, & Cook, 2002), and temporal and rhythmic information obtained from a beat (Tzanetakis, Jones, & McNally, 2007).).

II. Literature Survey

The domain of song recognition in music information retrieval (MIR) has been greatly advanced by the combination of machine intelligence and digital signal processing (DSP). The purpose of this review of the literature is to investigate current approaches, developments, and methods for identifying Indian music using DSP inside a machine intelligence framework.

Numerous studies on this subject suggested using RNN to categorize different musical genres. Networks designed for sequential data are the definition of this word (Yu et al., 2020). Unlike other Neural Network (NN) approaches, RNNs use connections created in a cycle to provide time-related context-based information for decision-making (Yu et al., 2020). The activations from one temporal phase are transferred to another by the connections (Yu et al., 2020). Long-term dependencies were outside the scope of the ordinary RNN structure due to potential vanishing gradient problems. Thus, in order to create a new connection state from successfully updated current activations, Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) were proposed (Yu et al., 2020).

In order for CNN to function, it needs to receive several spectrograms from the audio files as inputs and extract the patterns from them into a 2D convolutional layer using the proper filter and kernel sizes (Rafi et al., 2021). Lau suggested implementing the Convolutional Neural Network (CNN) model using a preprocessed GTZAN dataset. For every song in the collection, a Mel-Frequency Cepstrum Coefficient (MFCC) spectrogram was extracted (Lau & Ajoodha, 2021).

The Short-term Fourier Transform (STFT) spectrograms, which are made up of different sequences of spectrogram vectors across time, are the inputs used by Yu et al. to develop the CNN method (Yu et al., 2020). The GTZAN and Extended Ballroom databases were referenced in their paper. The data size set was increased by 18 times for each genre label when Yu et al. proceeded to extract each song from both datasets into 18 smaller parts in 3 seconds with a 50% overlap (Yu et al., 2020).

The concept of creating a 2D Convolutional Neural Network (CNN) was conceived by Athulya and Sindhu. They used the Librosa software and the GTZAN dataset to extract the audio recordings into several kinds of spectrograms. The binary inputs of a 2D CNN model were derived from those spectrograms using the Keras package. TensorFlow library was also used to generate the layers (Athulya et al., 2021). Convolutional neural networks (CNNs) were also suggested to be trained on one dimensional (1D) CNNs to identify different musical genres. The goal of Allamy and Koerich's 1D Resnet model was to keep the model from degrading and from having problems with vanishing gradients by substituting residual blocks for convolutional layers (Allamy & Koerich, 2021).

According to Jawaherlal Nehru, Jothilakshmi, and Nadu (2018), a Deep Neural Network (DNN) is used in the system's development to identify the genres. The properties of the music are represented by attributes called Mel Frequency Cepstral Coefficients (MFCC). MIR datasets are used to evaluate the system. With a greater classification accuracy of 97.8%, the suggested system was observed (Jawaherlalnehru et al., 2018).

III. Methodological Aspects

The process of training a machine learning system to predict the genre of a given music track is known as music genre categorization. Research on music genre classification is ongoing, with several methods being developed and tried to improve classification precision. In this work, the MGC goal is accomplished by the application of CNN, a deep learning approach. The process of training a machine learning system to predict the genre of a given music track is known as music genre categorization. Research on music genre classification is ongoing, with several methods being developed and tried to improve classification precision. In this work, the MGC goal is accomplished by the application of CNN, a deep learning approach.

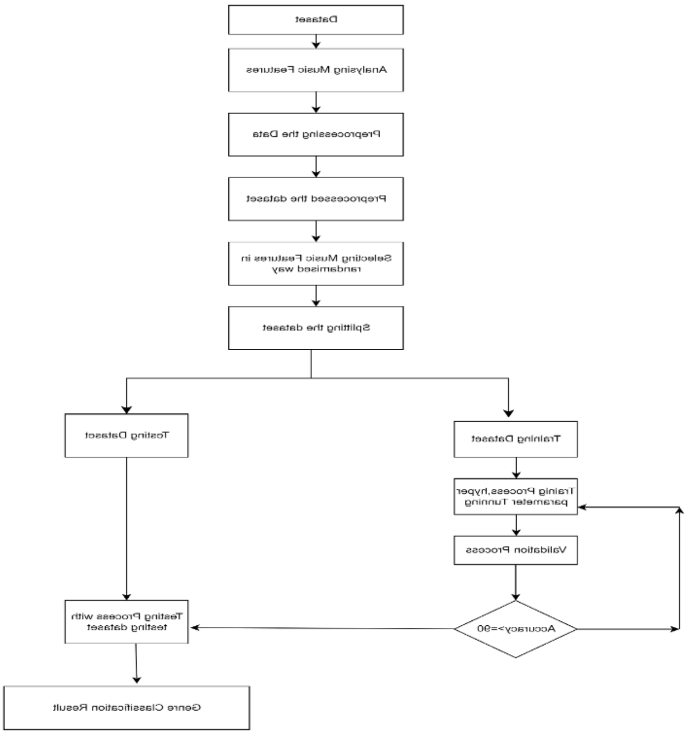


Fig 1. System Architecture

Dataset:

The dataset containing the original genres is a collection of 10 genres with 10 audio files each, all of which are 30 seconds long. Each audio file is represented by the original image. Neural networks are one tool used in data classification. To make this possible, the audio files were transformed to Mel spectra, since Neural Networks, such as CNN, usually employ some kind of picture representation. And lastly, two CSV files with the audio file's properties. From the audio recording, a file can be extracted including a mean (30 seconds) for each song and a calculated variance depending on multiple features. Although the songs in the other file were previously divided into a 3-second audio file, the structure of the other file is similar (this boosts the quantity of data we feed into our classification models by 10x). The genres are Bhajan, Ghazal, Thumri, Najrulgeeti, Rabindra Sangeet, Folk, Tandav, Thumri, Bollywood songs, Kawali.

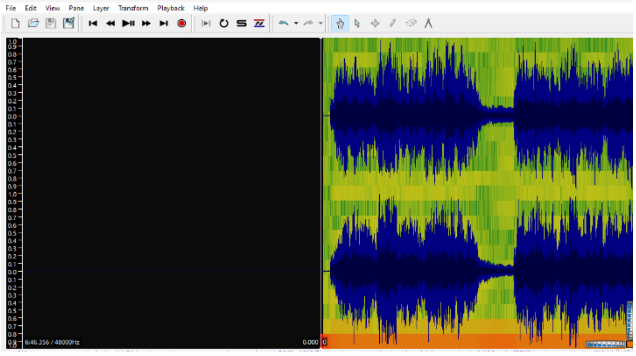


Fig 2. Spectrogram of Devotional song

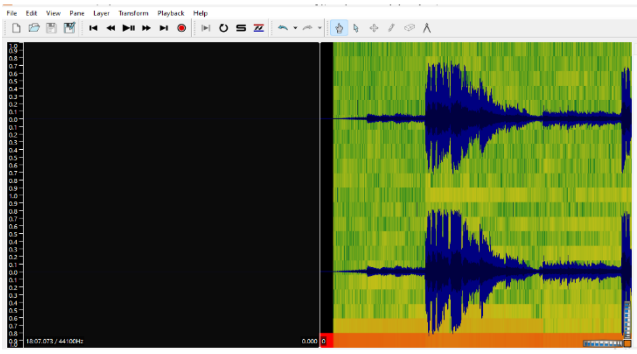


Fig 3. Spectrogram of Bhajan

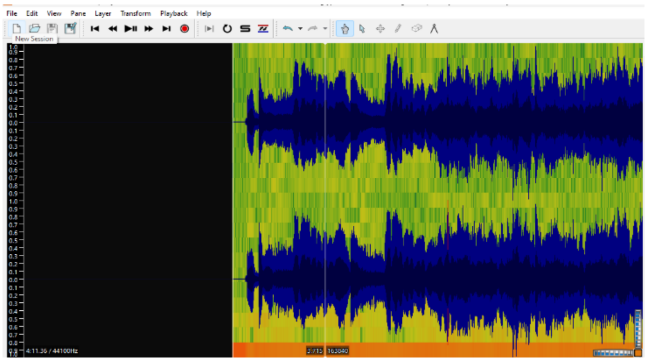


Fig 4. Spectrogram of Najrul Geeti

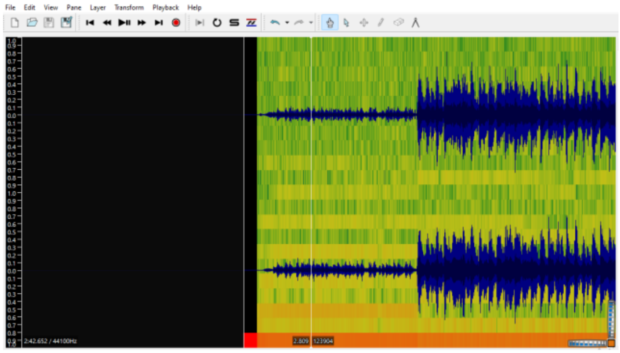


Fig 5. Spectrogram of Rabindra Sangeet

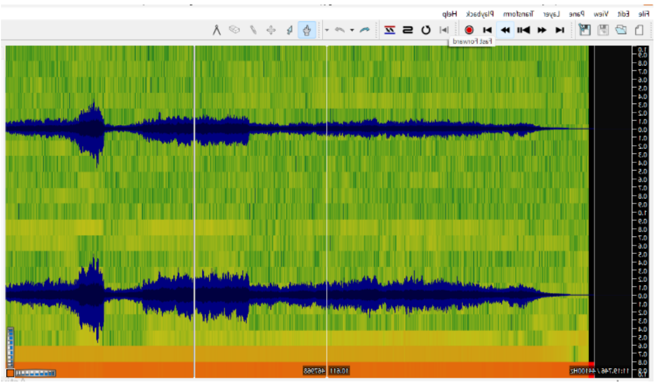


Fig 6. Spectrogram of Ghazal

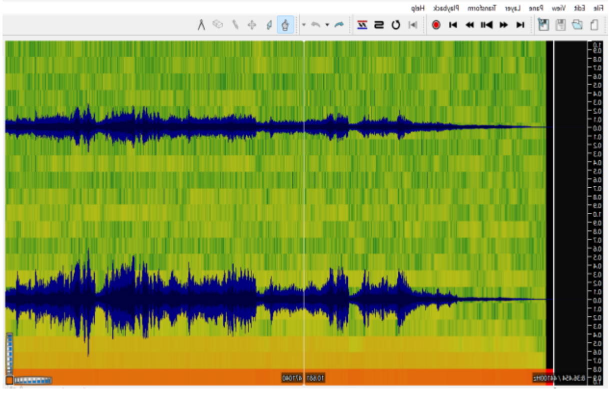


Fig 7. Spectrogram of Thumri

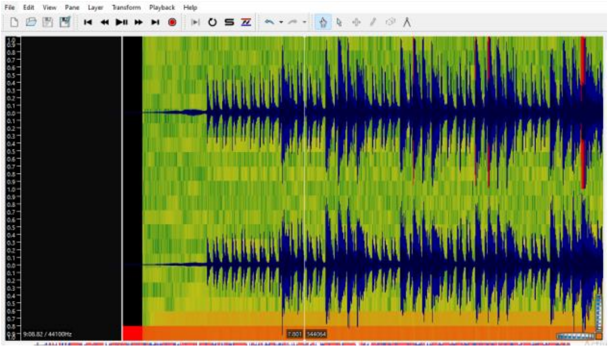


Fig 8 Spectrogram of Stotra

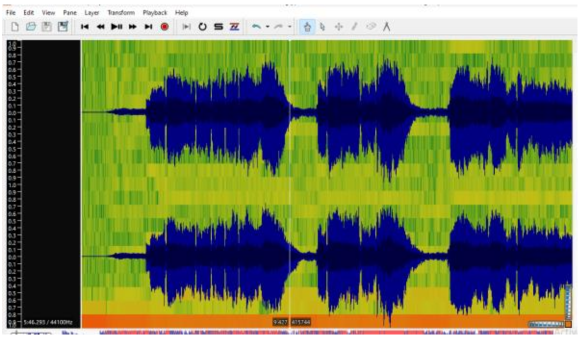


Fig 9. Spectrogram of Folk song

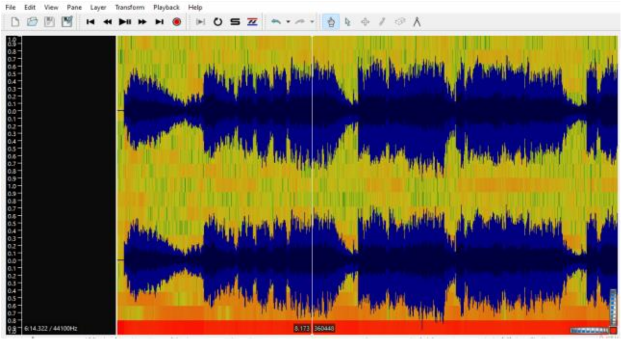


Fig 10. Spectrogram of Hindi Light Music

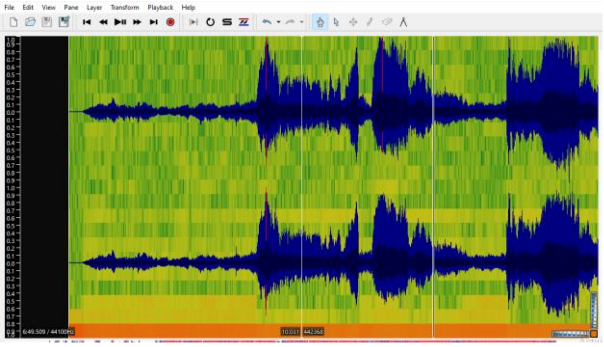
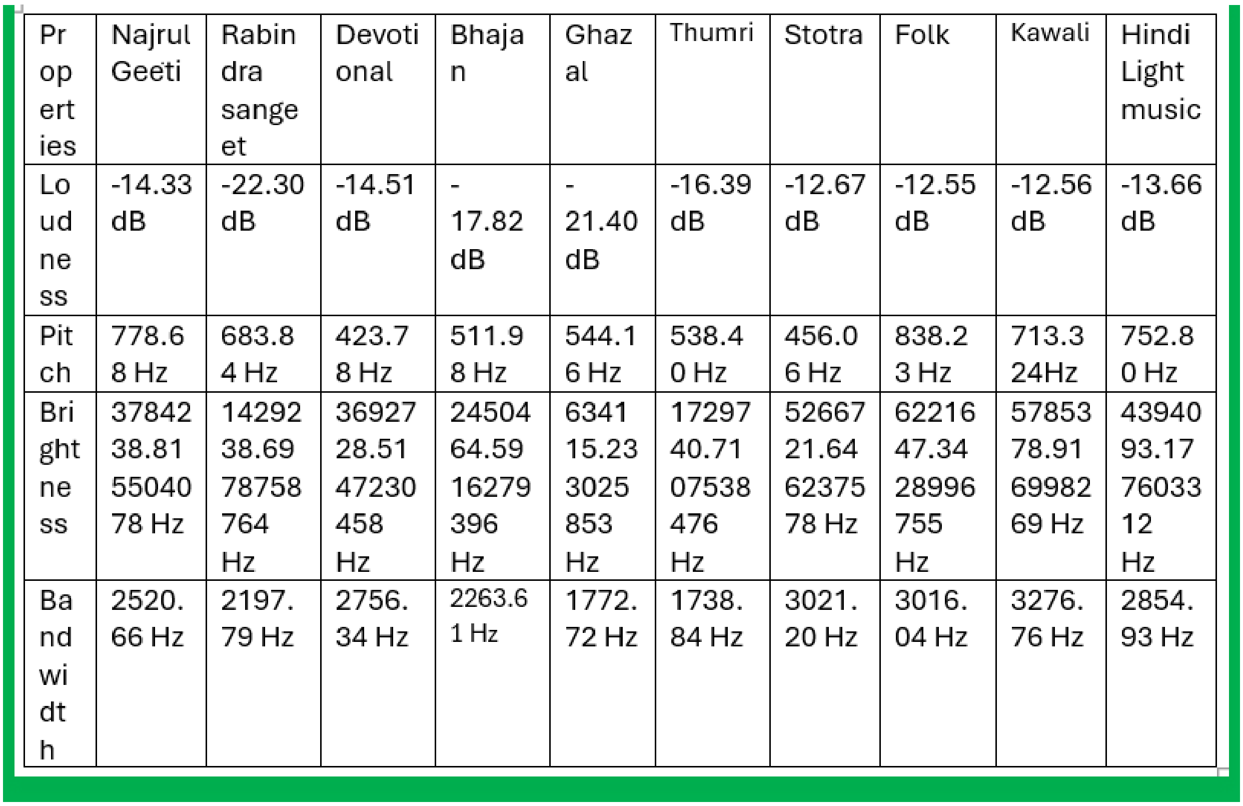


Fig 11. Spectrogram of Kawali



Properties	Najrul Geeti	Rabindra sangeet	Devotional	Bhajan	Ghazal	Thumri	Stotra	Folk	Kawali	Hindi Light music
Loudness	-14.33 dB	-22.30 dB	-14.51 dB	-17.82 dB	-21.40 dB	-16.39 dB	-12.67 dB	-12.55 dB	-12.56 dB	-13.66 dB
Pitch	778.68 Hz	683.84 Hz	423.78 Hz	511.98 Hz	544.16 Hz	538.40 Hz	456.06 Hz	838.23 Hz	713.324Hz	752.80 Hz
Brightness	3784238.815504078 Hz	1429238.6978758764 Hz	3692728.5147230458 Hz	2450464.5916279396 Hz	634115.233025853 Hz	1729740.7107538476 Hz	5266721.646237578 Hz	6221647.3428996755 Hz	5785378.916998269 Hz	4394093.177603312 Hz
Bandwidth	2520.66 Hz	2197.79 Hz	2756.34 Hz	2263.61 Hz	1772.72 Hz	1738.84 Hz	3021.20 Hz	3016.04 Hz	3276.76 Hz	2854.93 Hz

Fig. 12 Complete Data Analysis Result

Preprocessing:

Each music signal was processed using Librosa software, transforming the waveform into a mel-spectrogram with a temporal window of 23 ms. A logarithmic scale was then applied to the mel-spectrogram to normalize values across different mel scales. This technique, influenced by biological principles, provides a clearer understanding of the PCA-whitening process (Dong,2018). The dataset was split into 80% for training and 20% for testing.

Musical Features:

Feature extraction is a critical aspect of the classification system, especially in music genre classification. Accurate classification is only possible when music can be sufficiently characterized by its features. Musical characteristics are generally divided into two categories. The first is based on sensory attributes perceived by humans, such as note, pitch, and velocity. The second approach, as outlined (Lu,2019), classifies music features into short-term and long-term based on their duration.

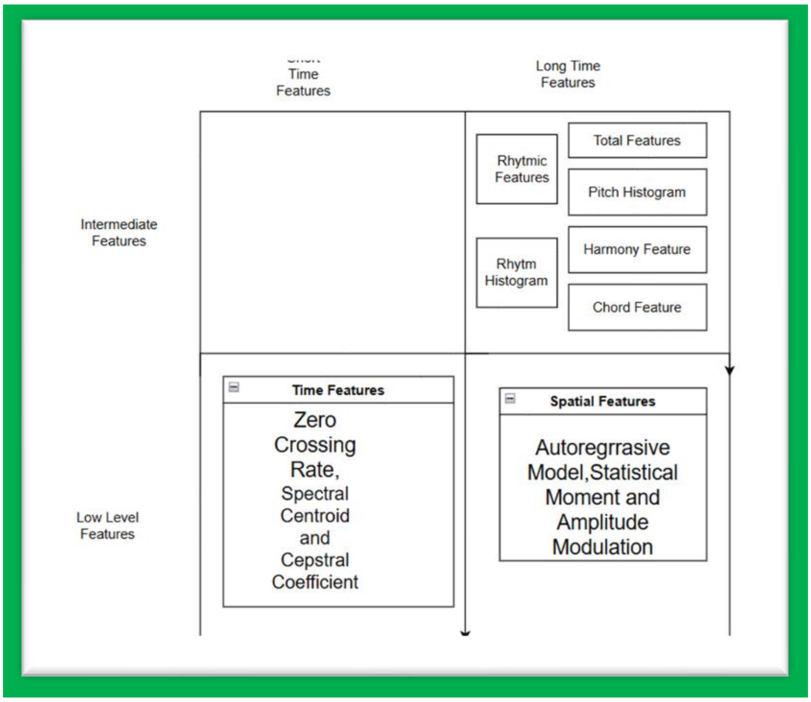


Fig 13. Classification of Music Features

Training and Testing:

10,0 mel-spectrogram-converted music files are split equally into testing, validation, and training sets in a 5:2:3 ratio. The training process is as follows:

a) Choosing a portion of the training set's tracks.

b) Select all the chosen tunes into 3-second continuous chunks, then randomly sample a starting point.

c) Use the back-propagation technique to calculate the gradients, using the segments as input and the original music labels as the target genres.

d) Use the gradients to update the weights.

e) Continue the process until the cross-validation data set's categorization accuracy stops improving. Every piece of music (mel-spectrogram) is split into 3-second segments with 50% overlap for testing purposes. Next, for every part, the trained neural network predicts the likelihood of each genre. For every piece of music, the genre predicted is the one with the highest averaged probability.

IV. CONVOLUTIONAL NEURAL NETWORK MODEL

A convolutional neural network, or CNN, is composed of one or more completely linked layers (like a standard neural multilayer network) after one or more dynamic layers. In this phase, the inserted picture, which has the size image_ width x image_height, is passed through a matrix filter (let's say 3x3). The element value is obtained by first applying the filter to the image matrix, then computing the intelligent repetition of the element between the filter and the picture region, and finally summarizing the results. There are four convolutional layers in the model. Convolutional filters, ReLU activation functions, and mass integration layers are the components that make up each layer. Before entering the neural network, there are two layers: a stop layer and a flat layer. The image tensor is transformed into a vector via the flat layer. An input for a neural network is this vector. Applying a stopping layer helps to avoid congestion. The neural network consists of an outward layer with nodes equal to the number of classes to be distinguished and a dense layer with 512 nodes.

V. CONCLUSION

Machine intelligence and digital signal processing techniques together have made it possible to create reliable and effective Indian music recognition systems. To solve lingering issues and investigate fresh approaches to improving the precision, effectiveness, and suitability of these systems in various real-world contexts, more research is necessary. This project proposes a neural network-based application for music genre classification. After considering several various audio feature extraction methods, it was determined that MFCC would work best in this situation. We used CNN and KNN techniques to classify our dataset and train our model. This research presents a neural network-based system for categorising music genres. One kind of music information retrieval (MIR) activity is music classification, when labels are applied to musical components like instrumentation, mood, and genre. The Python-based librosa package helps with feature extraction and, as a result, provides appropriate training settings for networks. Therefore, it seems that our approach has potential for classifying a sizable music database into the relevant genre.

References

1. Allamy, S., & Koerich, A. L. (2021). 1D CNN architectures for music genre classification. arXiv Preprint arXiv:2105.07302.
2. Athulya, K. M., et al. (2021). Deep learning-based music genre classification using spectrogram.
3. Burred, J., & Lerch, A. (2004). Hierarchical automatic audio signal classification. Journal of the Audio Engineering Society, 52(7/8), 724-739.
4. Carey, M., Parris, E., & Lloyd-Thomas, H. (1999). A comparison of features for speech/music discrimination. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 1437-1440). Phoenix, AZ: IEEE. https://doi.org/10.1109/ICASSP.1999.758084
5. Dannenberg, R. B., Thorn, B., & Watson, D. (1997). A machine learning approach to musical style recognition. In Proceedings of the International Computer Music Conference (ICMC) (pp. 344-347). Thessaloniki, Greece: ICMA.
6. Dixon, S., Pampalk, E., & Widmer, G. (2003). Classification of dance music by periodicity patterns. In Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR) (pp. 159-165).
7. Dong, M. (2018). Convolutional neural network achieves human-level accuracy in music genre classification. arXiv Preprint arXiv:1802.09697v1.
8. Foote, J. (1997). A similarity measure for automatic audio classification. In Proceedings of the 14th National Conference on Artificial Intelligence (pp. 275-280). Providence, RI: AAAI.
9. Jawaherlal Nehru, G., Jothilakshmi, S., & Nadu, T. (2018). Music genre classification using deep neural networks. International Journal of Scientific Research in Science, Engineering and Technology, 4(4), 935-940.
10. Lambrou, T., Kudumakis, P., Speller, R., Sandler, M. B., & Linney, A. (1998). Classification of audio signals using statistical features on time and wavelet transform domains. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Vol. 6, pp. 3621-3624). Seattle, WA: IEEE. https://doi.org/10.1109/ICASSP.1998.679665
11. Lau, D. S., & Ajoodha, R. (2021). Music genre classification: A comparative study between deep learning and traditional machine learning approaches. In Sixth International Congress on Information and Communication Technology (6th ICICT) (pp. 1–8). London, UK: Springer.
12. Li, S. Z. (2000). Content-based audio classification and retrieval using the nearest feature line method. IEEE Transactions on Speech and Audio Processing, 8(5), 619-625. https://doi.org/10.1109/89.861383
13. Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene classification. Journal of VLSI Signal Processing, 20(3), 343-348. https://doi.org/10.1109/MMSP.1997.602659
14. Lu, H. (2018). Music genre classification based on convolutional neural network. Electronic Measurement Technology, 42(21), 154-156.
15. McKinney, M., & Breebart, J. (2003). Features for audio and music classification. In Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR) (pp. 151-156). Baltimore, MD: ISMIR.
16. Quazi Ghulam Rafi, M., Noman, M., Sadia Zahin Prodhan, S., Alam, S., & Nandi, D. (2021). Comparative analysis of three improved deep learning architectures for music genre classification.
17. Saunders, J. (1996). Real-time discrimination of broadcast speech/music. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 993-996). Atlanta, GA: IEEE. https://doi.org/10.1109/ICASSP.1996.543290
18. Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 1331-1334). Munich, Germany: IEEE. https://doi.org/10.1109/ICASSP.1997.596192
19. Soltau, H., Schulz, T., Westphal, M., & Waibel, A. (1998). Recognition of music types. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 1137-1140). Seattle, WA: IEEE. https://doi.org/10.1109/ICASSP.1998.675470
20. Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293-302. https://doi.org/10.1109/TSA.2002.800560
21. Tzanetakis, G., Ermolinskyi, A., & Cook, P. (2002). Pitch histograms in audio and symbolic music information retrieval. In Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR) (pp. 31-38). Paris, France: ISMIR.

22. Tzanetakis, G., Jones, R., & McNally, K. (2007). Stereo panning features for classifying recording production style. In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR) (pp. 441-444). Vienna, Austria: ISMIR.
23. Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1996). Content-based classification, search, and retrieval of audio. IEEE Multimedia, 3(3), 27-36. https://doi.org/10.1109/93.556537
24. Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., & Feng, L. (2020). Deep attention based music genre classification. Neurocomputing, 372, 84–91. https://doi.org/10.1016/j.neucom.2019.08.078
25. Zhang, T. (1998). Hierarchical system for content-based audio classification and retrieval. Proceedings of SPIE, 398-409. https://doi.org/10.1117/12.325832