2024 5th International Conference for Emerging Technology (INCET)
Karnataka, India. May 24-26, 2024

Integrating Western and Indian Music for Genre Recognition: A Deep Learning Approach

Jayapriya A N
School of Electronics and Communication Engineering
KLE Technological University
Hubli, India
01fe21bec134@kletech.ac.in

A S V Dheeraj
School of Electronics and Communication Engineering
KLE Technological University
Hubli, India
01fe21bec161@kletech.ac.in

Shreya S Nadgir
School of Electronics and Communication Engineering
KLE Technological University
Hubli, India
01fe21bec127@kletech.ac.in

Satish Chikkamath
School of Electronics and Communication Engineering
KLE Technological University
Hubli, India
chikkamath@kletech.ac.in

Nirmala S R
School of Electronics and Communication Engineering
KLE Technological University
Hubli, India
nirmala.s@kletech.ac.in

Suneeta V Budihal
School of Electronics and Communication Engineering
KLE Technological University
Hubli, India
suneeta_vb@kletech.ac.in

2024 5th International Conference for Emerging Technology (INCET) | 979-8-3503-6115-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/INCET61516.2024.10593469

***Abstract*— Music genre recognition (MGR) is crucial for user preference and exploration and is often prioritised over other recommendation systems. Unlike previous studies that predominantly focused on Western or Indian music datasets, our paper addresses this notable gap in existing research by proposing a novel music genre classification methodology that uniquely integrates both Western and Indian music datasets. Our methodology, employing Convolutional Neural Network (CNN) and Mel-Spectrogram, achieves notable training accuracy and showcases promising results in automated music genre classification.**

Keywords*— *Convolutional Neural Network (CNN), Mel-Spectrograms, Music Genre Recognition.

I. INTRODUCTION

Genres are labels created by humans to differentiate between different styles of music [1]. The development of the internet has revolutionised the world by giving access to multimedia content like music. Numerous music streaming platforms, like Spotify and Wynk, have millions of songs and a lot of content consumers. Managing and accessing vast amounts of data is a significant challenge for both administrators and users [5][6][9]. Hence, the importance of creating music titles arises. In this case, music genres play a crucial role in creating solutions. They are also commonly used to organize music collections and represent intercultural connections.

Musical genres summarise and combine different music pieces' common properties and characteristics [13]. Distinguishing the music based on resemblance is called the music genre, meaning how similar a music is to the other audio belonging to the same genre [2 7 3 4]. Recognising the genre of a song based on its sound is a tidal yet vital task, keeping in mind that music breaks language barriers and can attract listeners. This detailed analysis of elements like rhythm, instruments, harmonic context, and tempo precisely categorises music into genres, which serves as valuable metadata for organizing music data. Music Information Retrieval (MIR) focuses on computational techniques to extract meaningful data from music in areas like instrument identification, artist recognition, emotion detection, plagiarism detection, and music genre categorization [5][6][9]. It targets the classification of music by genre, a method people use to explore the music of their preferences, often prioritising it over artist similarity, emotion, or other recommendation systems [8][9].

Data sets used in MGC (Music Genre Recognition) have no particular sequence or order. Moreover, the existing databases primarily focus on popular genres such as Classical, Electronic, Jazz, and many other Western-oriented music genres. Exploring and incorporating other music genres, particularly from India, is essential to enhance this automatic recognition process and extend it to other genres. This study aims to contribute by studying different genres but focuses more on Indian genres like Sufi, Classical, Folk, Indian Bollywood Pop, etc [13]. Data sets play a vital role in enhancing the efficiency of any model. Available data sets, like GTZAN, are the most suggested [12]. Other data sets, like Kaggle, are also used for Indian music datasets. The choice of music generally varies from person to person depending on personal tastes, location, age group, and various factors.

Categorising music based on its style, cultural impact, form, or genre is commonly used [3]. Since the early 2000s, traditional audio classification techniques like hidden Markov models (HMM), Support Vector Machine (SVM), K-nearest algorithms (KNN), and artificial Neural Networks (ANN) have been largely used for building models [16]. According to the study it has been observed that CNN-based models have shown high accuracy in the field of music genre classification, outperforming traditional machine learning algorithms [15]. Introducing the Deep Learning model is often more effective than traditional audio classification. Its ability to learn hierarchical features, end-to-end learning, large data handling, adaptability, and large data handling make it often more effective than traditional audio classification methods [16]. To understand a complex signal or sound, we need to perform certain operations, such as the Fourier Transform. Learning techniques represent a promising path forward in this dynamic field, where the harmony of data and algorithms strives to define the symphony of music genres converted to an MFCC spectrogram, which is used for audio sound classification. MFCC spectrogram obtained is provided as an input for Deep Learning Models with SoftMax and ReLu activation functions [10].

II. Literature Survey

In a research paper [9], Digital Signal Processing is used to analyse audio signals. Features related to rhythm and timbre are extracted via time and frequency domain methods. Fourier Transform (FT) such as DFT, FFT, and STFT are used in this process. Once the features are extracted, they use the metadata of the music track, which contains information like artist name, song name, label, genre, etc. After extraction, these features undergo normalisation and selection to ensure they are in the best form of classification. In order to improve music accuracy, Neural Networks, particularly CNN and RNN-LSTM models, and various other classifiers, including SVM, CNN, and DNN are used. Additionally, efforts are made to address the issue of overfitting in these models [14 8]. The paper outlines the extraction of time domain and frequency domain features using methods like statistical descriptors, zero-crossing rate, root mean square energy, and more. The paper examines various classification models, such as Random Forest Logistic Regression, Support Vector Machines, K-Nearest Neighbours, and XGBoost. It assesses the performance of these models based on parameters like accuracy, F1-score, and ROC AUC [3].

The provided technical paper outlines a research methodology for the classification of music genres using a dataset and a convolutional neural network (CNN) model. Wave plots are displayed to highlight the unique characteristics of each genre based on amplitude and frequency. All audio samples in the dataset are considered input, standardised at 30 seconds, and in WAV format. This paper focuses on feature extraction, particularly amplitude features, to capture relevant information from audio samples. A ten-layer CNN architecture is employed for classification, incorporating eight Dense layers and two dropout layers to address overfitting. Performance evaluation is based on recognition rate, precision, F1-score, recall, support and confusion matrix. The GTZAN dataset is utilised, which includes ten genres and demonstrates impressive accuracy (98.30). In the research paper [10], to achieve music genre classification, the study employs supervised machine learning, especially the K Nearest Algorithm and Random Forest. The use of these algorithms is due to their efficiency in handling pre-labelled training data [14]. In the research paper [12], the data set consists of 816 instances, with the audio wav-file serving as the independent variable and the Music Genre Classifications as the dependent variable.

A research paper [13], which assessed the efficiency of various machine learning algorithms using MFCCs as attributes and selected genres from databases like GTZAN [8]. For differentiating between Metal and Jazz they found Neural Networks to be more efficient and superior.

A unique multi-class classification approach was employed using the LMD (Latin Music Database), where music's space-time segments were used for binary classifications. Other datasets include the million-song dataset (MSD), MTG-Jamendo dataset, YouTube-8M dataset, and MuMu Dataset [16]. These results were then merged to finalise the genre label, proving more accurate than using the fixed song segments. Comparison with other studies like MGR (Music Genre Classification) due to the difference in data extraction methodologies. The activation function serves as a simple function that takes in the output of a node. The Adam optimizer operates similarly to a variable learning rate descent, incorporates batching [8], and is used as an estimator for developing deep models.

The CNN classifier is implemented in two manners: firstly, a custom sequential CNN model and secondly a pre-trained convent called Mobile Net [1]. Other algorithms, including Resnet34, Resnet50, and AlexNet also be used for the same. Another approach that can be utilized is the parallel recurrent convolutional neural (PRCNN), which is highlighted in the paper, employing a combination of Bi-RNN and CNN for spatial feature extraction and classification, showing promising results in tasks such as music genre classification [16]. Additionally, in a research paper [16], the focus is on audio classification using a range of deep learning models. They converted 1D audio signals to 2D spectrograms through feature extraction techniques like STFT, Mel- spectrograms and MFCCs. CNN is mostly preferred for spatial feature extraction, RNN is preferred for time dependency tasks, transformers for audio representation, and hybrid models combining these approaches were employed.

The research suggests a promising future for deep learning in audio classification, mostly focused on transformer-based and hybrid models. Traditional machine learning algorithms like Logistic Regression (LR), Gaussian Discriminant Analysis (GDA), Random Forest (RF), and Support Vector Machine (SVM) are used alongside a fine-tuned VGG-16 CNN [15][14]. The CNN-based approach achieves an accuracy of 88.5. The raw audio data is converted into an image-like format. Discrete Fourier Transform (DFT) is applied to extract frequency features. We prefer the Short-Time Fourier Transform (STFT) for creating spectrograms to obtain greater frequency representation. Mel spectrogram is generated to represent the human-scale hearing range. Overfitting is observed, but it is mitigated using ridge regression [11]. The model's performance is evaluated by parameters such as AUC, accuracy, and F-score. Confusion matrices are created to assess classification results. An ensemble classifier combining both CNNs and the traditional approach with hand-crafted features achieved an AUC value of 0.894 [14]. The paper explores genre-based music classification across three distinct categories: audio-based features, image-based features, and features derived from modalities [16].

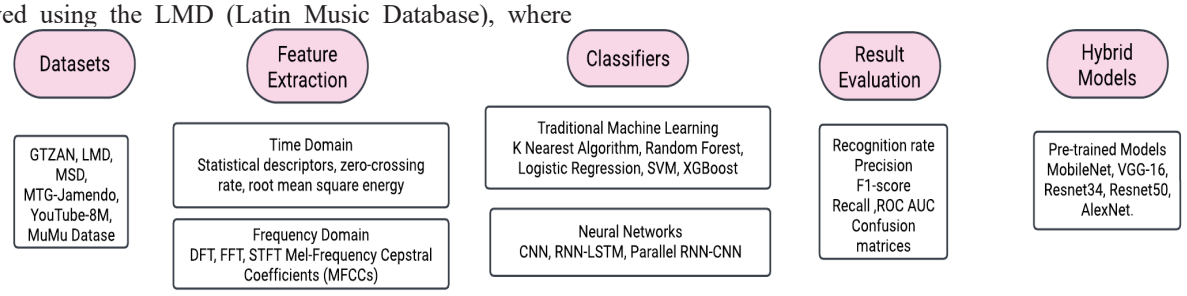


Fig. 1. Flowchart showing steps involved in the training model to identify Music genre

III. METHODOLOGY

A. Datasets

The datasets used are labelled music datasets taken from GTZAN and Kaggle. There are 13 different music classes containing 100 samples for every class. These classes include Semi Classical, Hip-hop, Ghazal, Disco, Country, Carnatic, Blues, Sufi, Rock, Reggae, Pop, Metal, and Jazz all of which have a duration of 30 seconds and a frequency of 44100 Hz.

B. Proposed Method

The first step involves the organisation of a labelled music dataset, which involves categorising audio files by genre and placing them in appropriately named folders.

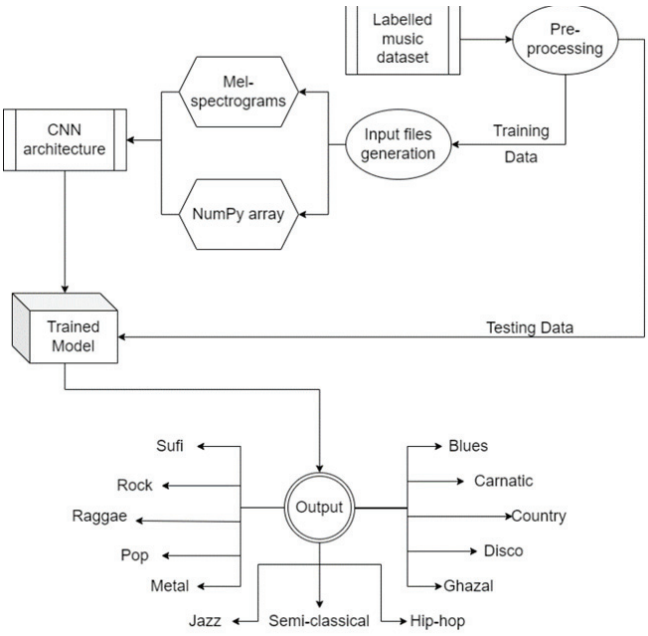


Fig. 2. Proposed framework for MGR

In the pre-processing phase, we enhance computational efficiency by standardising the audio samples. This involves adjusting the sample rate to ensure uniformity across all the samples to 10-second files. Additionally, we truncate the original 30-second audio intervals to streamline processing and focus on relevant segments.

The dataset is divided into three subsets for training, validation, and testing. The division is conducted utilising the train, test, split method from the Skilearn library, with 80% of the data allocated for training, 10% for validation, and a further 10% for testing.

Our approach employs Mel-spectrograms to generate the spectral characteristics of the processed audio signal. The input data (X) consists of a collection of 2D NumPy arrays, each representing the Mel-spectrogram, while the output labels (y) are integers signifying the genre index for each corresponding Mel-spectrogram in X. The extracted features are utilized as input for the CNN model, which is initially trained and subsequently employed to recognize the audio signals and their corresponding classes.

C. The architecture of the CNN model

The CNN model uses 7 layers, which includes one input layer, 3 convolution layers, namely conv1, conv2, and conv3, one flatten layer, and two dense layers namely dense1 and dense2.

1) Input layer:

- Accepts a 2D array representing Mel-Spectrogram, capturing audio features.
- Additionally, it takes an integer representing the genre label associated with the input
- Also, input images are resized to 128 x 128 pixel

2) Convolution Layers

- Three sets of Conv2D + MaxPooling2D + Dropout layers.
- Each Conv2D layer has 64 filters with a 3x3 kernel and ReLU activation.
- MaxPooling2D layers with a pool size of (2, 2) for spatial down sampling.
- Dropout layers with dropout rates of 0.2 after each Conv2D layer.

3) Flatten Layer and Dense Layers:

- Flatten layer converts the 2D feature maps to a 1D vector.
- Dense layer with 128 units and ReLU activation.
- Dropout layer with a dropout rate of 0.4 after the dense layer.

4) Output Layer:

- Dense layer with 13 units (for 13 classes) and softmax activation for classification.

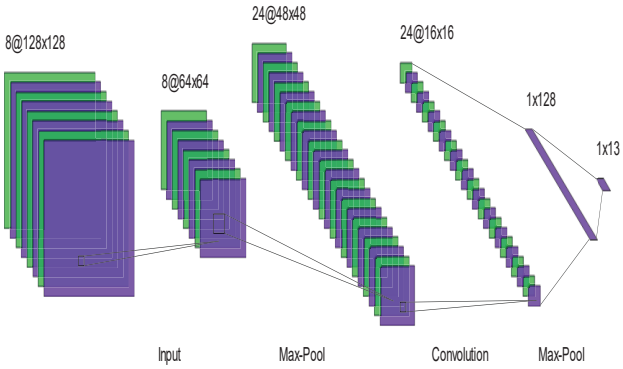
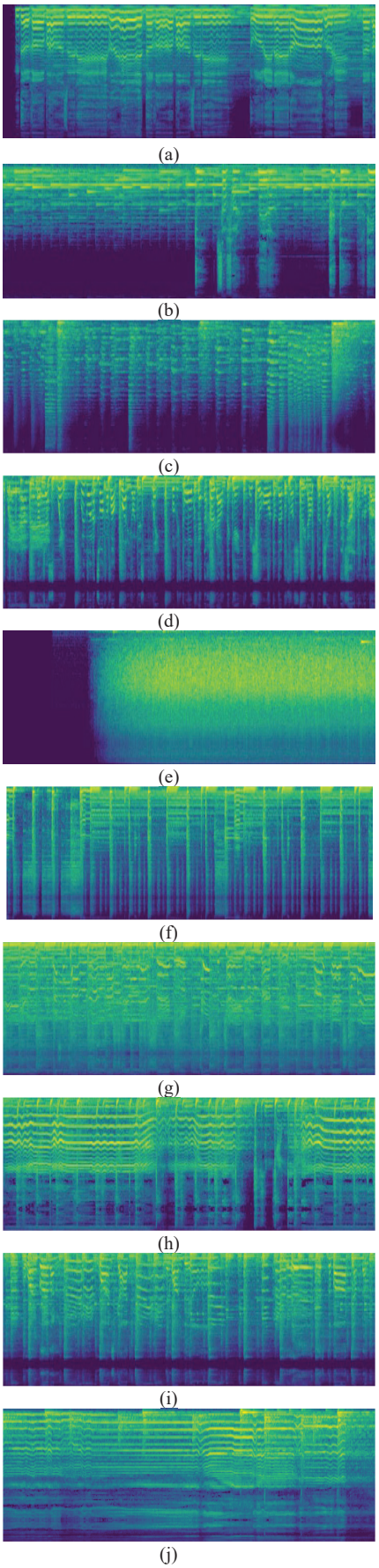


Fig. 3. CNN architecture for MGR model

IV. RESULTS AND OBSERVATIONS

Initially, we started training with 10 epochs for two genres and achieved an accuracy of 95%. We increased our genre to 13, trained the model for 10 epochs, and achieved an accuracy of 80%. As this was not sufficient, we increased the number of epochs to 20 and achieved an accuracy of 85.3%, and with the number of epochs increased to 50, we were able to achieve an accuracy of 88.46%.



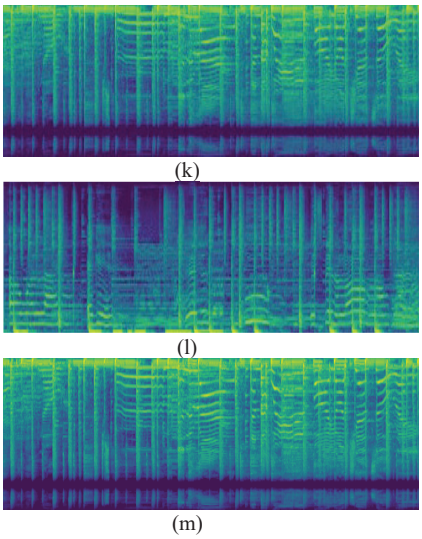


Fig. 4. a, b, c, d, e, f, g, h, i, j, k, l, m represents the spectrograms of the music classes Sufi, Jazz, Hip-hop, Gazal, Disco, Country, Carnatic, Pop, Blues, Semi-classical, Rock, Reggae and Metal respectively.

Spectrogram pictures help to illustrate and distinguish between different musical genres by offering a visual portrayal of the frequency content and temporal dynamics of audio signals. Spectrograms demonstrate the distinctive instrumental timbres, temporal structures, and frequency distributions found in music recordings by encapsulating the distinct spectral patterns linked to various genres. Researchers and machine learning algorithms can extract useful information for genre analysis and categorization from these visual representations. Spectrophotogram pictures provide insights into the many auditory landscapes of various music genres, from the rhythmic complexity of dance music to the melodic subtleties of classical compositions. This allows for a greater knowledge and investigation of musical styles and trends.

Time: A spectrogram's horizontal axis usually depicts time, with each point on the axis denoting a distinct point in time.

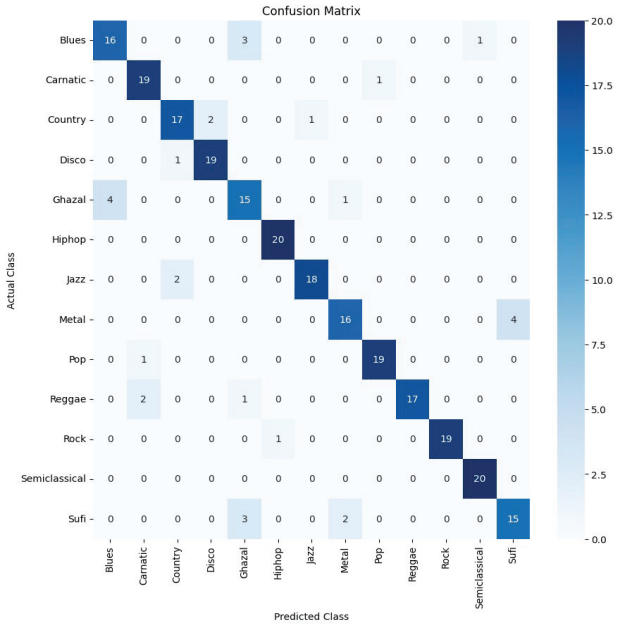


Fig. 5. Confusion matrix for predicted vs. actual class

Frequency: The range of frequencies contained in the signal is represented by the vertical axis, which stands for frequency. Higher frequencies are often shown at the top of the spectrogram, while lower frequencies are typically shown at the bottom.

Intensity or Amplitude: Colour or grayscale shading is used to indicate the signal's strength at each frequency and time interval. At a specific frequency and time, brighter or more vivid colours usually signify more signal energy or amplitude

In the above confusion matrix, each row represents the actual class of data points, and each column represents the predicted class. The diagonal cell of the matrix shows the number of data points that were correctly classified. The off-diagonal is incorrectly classified.

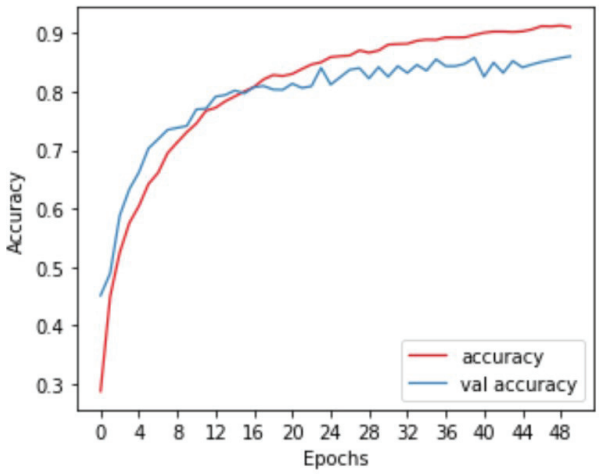


Fig. 6. Accuracy vs epoch graphs

V. CONCLUSION

The results indicate that the proposed MGR framework utilises a Mel-Spectrogram and a CNN model and is effective in recognising music genres. The optimisation steps, including increasing epochs and adding a convolution layer, significantly improved the model's accuracy. However, achieving high accuracy may be influenced by factors such as the diversity and complexity of music genres, data size, and model architecture.

In conclusion, the developed MGR framework demonstrates promising results in music genre recognition. Further refinement and exploration, including experiments with different model architectures and expanding the dataset, could enhance the model's performance for a broader range of music genres. The application of this framework extends to music recognition systems, content categorization, and other areas where accurate genre identification is essential.

Our research paper is an innovative approach to music genre classification, differentiating itself from previous studies that focused on singular Western or Indian music datasets. Unlike models trained exclusively on specific regional genres, our methodology leverages a unique integration of both Western and Indian music datasets, posing a challenge in direct comparison with models like Logistic Regression (LR), Gaussian Discriminant Analysis (GDA), Random Forest (RF), and Support Vector Machine (SVM) that are used alongside a fine-tuned VGG-16 CNN exclusively trained on a single type of data. This inclusive approach provides a comprehensive perspective on music genre classification, extending beyond the limitations of previous studies confined to specific cultural or regional contexts.

ACKNOWLEDGMENT

The authors would like to acknowledge KLE Technological University, Vidyanagar, Hubli, Karnataka for their necessary support

REFERENCES

[1] N. M R and S. Mohan B S,"Music Genre Classification using Spectrograms," 2020 International Conference on Power, Instrumentation, Control and Computing (PICC), Thrissur, India, 2020, pp. 1-5, doi: 10.1109/PICC51425.2020.9362364.

[2] S. Deepak and B. G. Prasad,"Music Classification based on Genre using LSTM," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 985-991, doi: 10.1109/ICIRCA48905.2020.9182850.

[3] V. Shah, A. Tandle, N. Sharma and V. Sheth,"Genre Based Music Classification using Machine Learning and Convolutional Neural Networks," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-8, doi: 10.1109/ICCCNT51525.2021.9579597.

[4] S. Pasrija, S. Sahu and S. Meena,"Audio Based Music Genre Classification using Convolutional Neural Networks Sequential Model," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavala, India, 2023, pp. 1-5, doi: 10.1109/I2CT57861.2023.10126446.

[5] J. Martins de Sousa, E. Torres Pereira and L. Ribeiro Veloso, "A robust music genre classification approach for global and regional music datasets evaluation," 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, China, 2016, pp. 109-113, doi: 10.1109/ICDSP.2016.7868526.

[6] S. Allamy and A. L. Koerich, "1D CNN Architectures for Music Genre Classification," 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 2021, pp. 01-07, doi: 10.1109/SSCI50451.2021.9659979.

[7] Y. Yuniar, D. P. Alamsyah and A. Herliana,"Classification of Indonesian Music Genres Using the Support Vector Machine Method," 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), Prapat, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICORIS56080.2022.10031473.

[8] R. Sharma and Nisha,"Classification of Music Genres using Neural Network," 2022 11th International Conference on System Modeling Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 142-147, doi: 10.1109/SMART55829.2022.10046811.

[9] T. Ozseven and B. E. ¨ Ozseven, "A Content Analysis of the Research Ap- ¨ proaches in Music Genre Recognition," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2022, pp. 1-13, doi: 10.1109/HORA55278.2022.9799935.

[10] M. S. Ahmed, M. Z. Mahmud and S. Akhter," Musical Genre Classification on the Marsyas Audio Data Using Convolution NN," 2020 23rd International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh, 2020, pp. 1-6, doi: 10.1109/ICCIT51783.2020.9392737.

[11] W. Suo,"Efficient Music Genre Classification with Deep Convolutional Neural Networks," 2022 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 2022, pp. 01-05, doi: 10.1109/DSIT55514.2022.9943952.

[12] M. P. V. N. Sai and S. Kalaiarasi,"Implementation of Music genre classification using Support Vector Clustering algorithm and KNN Classifier for improving accuracy," 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICONSTEM56934.2023.10142741.

[13] J. L. Concei¸c˜ao, R. de Freitas, B. Gadelha, J. G. Kienen, S. Anders and B. Cavalcante, "Applying supervised learning techniques to Brazilian music genre classification," 2020 XLVI Latin American Computing Conference (CLEI), Loja, Ecuador, 2020, pp. 102-107, doi: 10.1109/CLEI52000.2020.00019.

[14] Chikkamath, Satish, and S. R. Nirmala. "Melody generation using LSTM and BI-LSTM Network." In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, pp. 1-6. IEEE, 2021.

[15] Khudavand, Atiq Ahmed, Satish Chikkamath, S. R. Nirmala, and Nalini Iyer. "Music/Non-music Discrimination Using Convolutional Neural Networks." In *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 1*, pp. 17-28. Springer Singapore, 2021.

[16] Chikkamath, Satish, and S. R. Nirmala. "Music Detection Using Deep Learning with Tensorflow." In *ICDSMLA 2020: Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications*, pp. 283-291. Springer Singapore, 2022.