Music Genre Classfication

Atmakuri Rama Subramanyam
Department of CSE,
VIT-AP University, India
subramanyam.21bce9111@vitapstudent.ac.in

Thippulareddygari Jawahar Reddy
Department of CSE,
VIT-AP University, India
jawaharreddy.21bce9023@vitapstudent.ac.in

Koduru Hajarathaiah
Department of CSE,
VIT-AP University, India
hajarathaiah.k@vitap.ac.in

Mohammad Muneeb
Department of CSE,
VIT-AP University, India
muneeb.21bce9225@vitapstudent.ac.in

Marra Cheran Sre Josh
Department of CSE,
VIT-AP University, India
cheransre.21bce9537@vitapstudent.ac.in

***Abstract*—Music genre is a conventional category that identifies pieces of music as belonging to a shared tradition or set of conventions. In modern times, internet-based technologies have transformed the music industry, leading to the rise of online music streaming services. These platforms use music classification to recommend songs to users based on their listening patterns. Where we have used 10 distinct genres: like blues,classical,country,disco,hip-hop,jazz,metal,pop,reggae,and rock. In our work, we employed Convolutional Neural Networks (CNNs) to classify music using spectrogram images. The CNN classifies the music by analyzing the frequency patterns captured in the spectrograms. We used several neural networks like VGG16,VGG19,CNN,CRNN and CRNN with Attention mechanism. We optimized the models using ADAM optimizer and improve the accuracy.We have used large-scale music data to achieve the best output. Our goal is to evalute these models accurately and enhace their model to classify correctly.I have used metrics like accuracy, F1 score, precision, and recall to evaluate models.**

***Index Terms*—Deep learning algorithms, Spectogram Images, Convolutional Neural Networks.**

I. Introduction

The widespread use of the internet opened the door for significant advancements and modifications in the music business. A noteworthy illustration of these advancements is the widespread adoption of internet-based music streaming services. These platforms are primarily concerned with controlling author copyrights and classifying music in order to mine listening data, tag music, and make recommendations for music that will boost sales. In particular, people have access to millions of songs at all times and from any location. The number of ways that Convolutional Neural Networks (CNNs) are being used for audio classification has increased recently. Mel-Frequency Cepstrum Coefficients (MFCC) or log cepstrum are used as input in the majority of them. After the signal's Fourier transform, the logarithm operation is called a logcepstrum. The spectrogram is then obtained by performing the inverse Fourier transform. The axis of the spectrogram is logarithmic to conform to human perception because the cepstrum was originally used to measure seismic waves, and human perception of sound and seismic waves is similar to the cepstrum.[1] Predicting a music clip's genre from its auditory signal is the task of music genre categorization. Recent methods combine feature extraction and classification using Deep Neural Networks (DNNs), enabling the system to simultaneously learn pertinent characteristics and classify them. Using data like mel-spectrograms and raw audio signals, several DNN-related methods have been developed for automatic music labeling.[2][3] Support vector machines and neural networks two popular machine learning techniques have long been used for the second stage, the features that are extracted from the audio are usually designed rather than taught. By learning hierarchical features through machine learning algorithms; convolutional kernels, Convolutional Neural Networks (CNNs) have been widely employed for music classification tasks, including genre categorization and music tagging. In order to improve performance on music classification tasks, CNNs and Recurrent Neural Networks (RNNs) have recently been merged to construct Convolutional Recurrent Neural Networks (CRNNs), where CNNs extract features and RNNs represent sequential data like audio signals.[4] While the mechanism of learned filters in CNNs is evident for known target shapes, their functioning in tasks like mood detection or genre categorization remains uncertain. Despite reaching state-of-the-art performance in these tasks, researchers continue to investigate how CNNs learn and express subjective sensations connected to acoustical features.[5]

Machine learning (ML) techniques can quickly and accurately analyze large amounts of data, leading to more precise diagnoses and predictions.[6] [7] In our research, we focus exclusively on music genre classification without venturing into music recommendation. We use state-of-the-art neural network architectures, including CNN, CRNN, VGG16, VGG19, and CRNN with attention mechanisms, to enhance genre classification accuracy. Our approach aims to explore how these models perform in genre classification and improve the overall system's ability to classify music genres.

The main motivation for developing this project is the rise of online streaming platforms which has greatly speeded up the digital transformation of the music industry, making it crucial

to implement effective music genre classification for managing copyrights, improving user experience, and optimizing recommendations. While deep learning models like Convolutional Neural Networks (CNNs)[8] and Convolutional Recurrent Neural Networks (CRNNs) can learn complex patterns directly from audio data, traditional methods that rely on manually crafted features have their drawbacks. This study seeks to leverage advanced architectures, including CNN, CRNN, VGG16, VGG19, and attention mechanisms, to improve both accuracy and scalability in music genre classification, thus meeting the growing demands of the music industry.

In this research, various ML techniques will be employed to assess accuracy, precision, recall, and the F1 score. A comparative study will be conducted to evaluate the effectiveness of various supervised learning methods in breast cancer prediction.

Accuracy Score: A key metric in assessing machine learning model performance, measures the ratio of correct predictions to total predictions. It indicates how reliably the model predicts outcomes, with a higher score suggesting greater accuracy in achieving the desired output. [True Positives(TP), True Negatives(TN), False Positives(FP), False Negatives(FN)]

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

Precision Score: This is a performance metric used in machine learning (ML) to evaluate how well a model predicts the future. It determines the precision by dividing the number of true positives (TP) by the total number of false positives (FP) and true positives (TP). A greater precision score, which is typically expressed as a percentage, denotes superior model performance.

$$PS = \frac{TruePositives}{TruePositives + FalsePositives}$$

Recall Score: In machine learning, this statistic assesses how well a model can locate pertinent instances within a dataset. Recall is calculated by dividing the total number of relevant examples overall by the number of correctly identified relevant events. Superior model performance in correctly identifying pertinent events is indicated by a higher recall score.

$$RS = \frac{TruePositives}{TruePositives + FalseNegatives}$$

F1-Score: A critical measure in machine learning for assessing classification model performance, it's the harmonic mean of recall and precision. This metric considers both false positives and negatives, making it valuable for imbalanced class distribution or equal weighting of precision and recall. It's commonly used alongside accuracy, precision, and recall.

$$F1 = \frac{2 * PS * RS}{PS + RS}$$

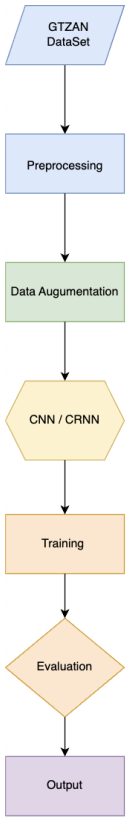

Fig. 1. Network Architecture

II. Data sets

We utilized the GTZAN dataset, a widely recognized benchmark for music genre classification tasks. The dataset comprises 1,000 audio files, evenly distributed across 10 distinct genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each audio file has a duration of 30 seconds and is recorded in Mono at a sampling rate of 22,050 Hz, with a 16-bit depth, and is stored in the .wav format. [2][9]This dataset provides a well-balanced representation of various musical genres, making it an ideal resource for evaluating the performance of classification algorithms.

We generated Mel spectrogram images from the audio files to serve as input data for our classification models. Mel spectrograms provide a time-frequency representation of audio, effectively capturing features relevant to distinguishing musical genres.[10] [11]Given the limited number of audio samples in the GTZAN dataset[12], we expanded the dataset using data augmentation techniques. This augmentation included transformations such as random horizontal flipping, small rotations, and scaling, which introduced variability while preserving the essential genre characteristics.[13] These augmented spectrogram images enhanced the model's robustness and ability to generalize, leading to improved performance in music genre classification.

Layer	Configuration	Output Dimensions
Input	Mel-spectrogram (96 × 1366 × 1)	96 × 1366 × 1
Conv Layer 1	Conv 3×3×128, stride=1, padding=1	96 × 1366 × 128
Max Pooling	MP (2, 4)	48 × 341 × 128
Conv Layer 2	Conv 3×3×256, stride=1, padding=1	48 × 341 × 256
Max Pooling	MP (2, 4)	24 × 85 × 256
Conv Layer 3	Conv 3×3×512, stride=1, padding=1	24 × 85 × 512
Max Pooling	MP (2, 4)	12 × 21 × 512
Conv Layer 4	Conv 3×3×1024, stride=1, padding=1	12 × 21 × 1024
Max Pooling	MP (3, 5)	4 × 4 × 1024
Conv Layer 5	Conv 3×3×2048, stride=1, padding=1	4 × 4 × 2048
Max Pooling	MP (4, 4)	1 × 1 × 2048
Fully Connected 1	Conv 1×1×1024	1 × 1 × 1024
Fully Connected 2	Conv 1×1×1024	1 × 1 × 1024
Output Layer	Conv 1×1×50, Sigmoid	1 × 1 × 50

TABLE I
LAYER CONFIGURATION AND OUTPUT DIMENSIONS

III. TRAINING AND TESTING

The preprocessing of spectrogram images begins with converting these images to grayscale, as the color channels in spectrograms often contain redundant information.[13] This conversion reduces the data's complexity, allowing the model to focus on essential features. Next, each image is resized to a uniform dimension, such as 128x128 pixels, to maintain consistency across inputs and ensure compatibility with input layer requirements. Additionally, normalization is applied to standardize pixel values, typically scaling them to a range of 0 to 1.[14]

- **CNN**: The CNN architecture features two convolutional layers with filters, each followed by ReLU activation and max pooling for down-sampling. The output is flattened and passed through a fully connected layer with units and a dropout layer to mitigate overfitting, concluding with another fully connected layer that generates the class logits.
- **VGG16/VGG19**: The VGG16 model is adapted to process grayscale spectrogram images by modifying its input layer for single-channel data and adjusting the output layer to match the dataset classes. Inputs are resized with normalization. An Adam optimizer with a learning rate and a StepLR scheduler guides the training process, with model performance evaluated on the test set.
- **CRNN**: The CRNN architecture integrates convolutional and recurrent neural networks to effectively capture both spatial and temporal information from spectrogram images. This model is particularly well-suited for audio classification tasks and is trained using cross-entropy loss with the Adam optimizer, potentially enhanced by data augmentation techniques for improved performance and robustness.
- **CRNN with Attention**: This model combines CNNs and bidirectional LSTMs for audio classification, utilizing an attention mechanism to enhance focus on important features. It processes audio data using Mel-Frequency Cepstral Coefficients (MFCC) and employs techniques like noise augmentation, Cross-Entropy loss, and Adam optimization with early stopping.

IV. RESULTS

Model	Accuracy	Precision	Recall	F1 Score
CNN	0.777171	0.7771	0.7712	0.7717
CRNN	0.593939	0.6065	0.5845	0.5874
VGG16	0.795454	0.7924	0.7924	0.7913
VGG19	0.865151	0.9436	0.9425	0.9425
CRNN(Attent)	0.50606	0.5162	0.506	0.4951

TABLE II
MODEL PERFORMANCE COMPARISON

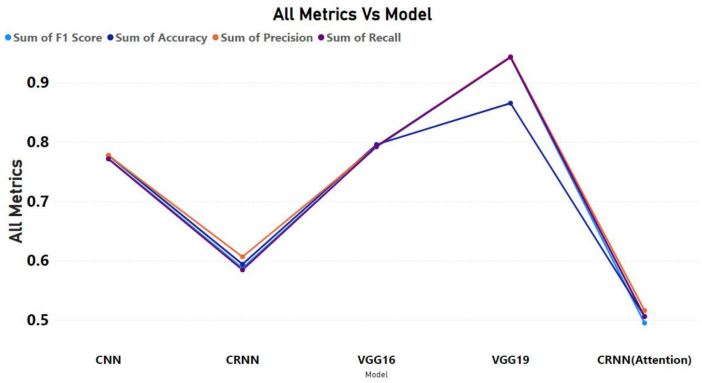

Fig. 2. Image showing All Metrics

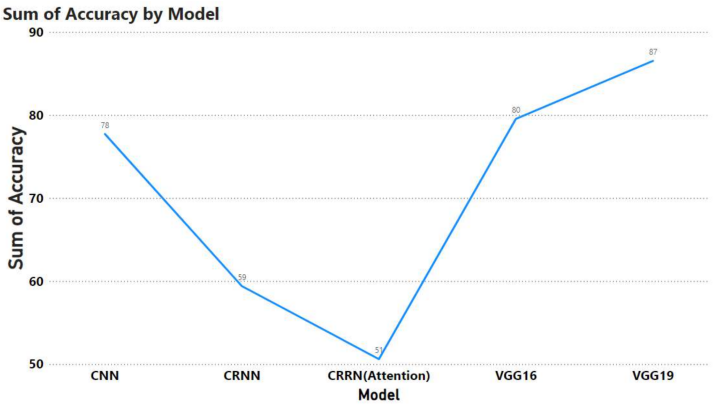

Fig. 3. Image showing Model Accuracy

The performance of five deep learning models—CNN, CRNN, VGG16, VGG19, and CRNN with Attention—was evaluated using four metrics: accuracy, precision, recall, and F1-score. Among

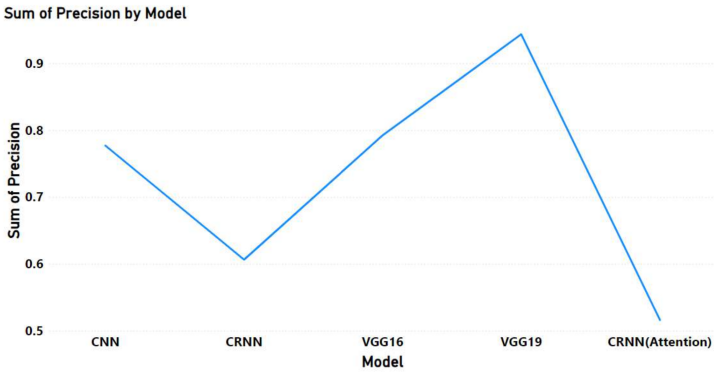

Fig. 4. Image showing Model Precision

Fig. 6. Image showing Model F1 score

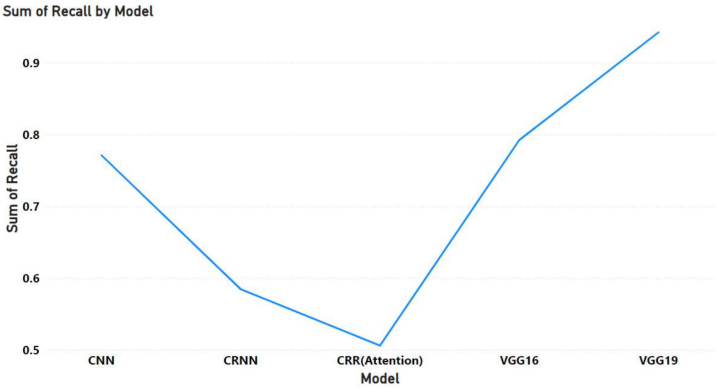

Fig. 5. Image showing Model Recall

these, VGG19 outperformed the others, achieving the highest accuracy (86.52%), precision (0.9436), recall (0.9425), and F1-score (0.9425), indicating its superiority in balancing correct predictions across both positive and negative classes. VGG16 followed closely with an accuracy of (79.55%) and consistent precision and recall values of (0.7924), making it the second-best model. CNN demonstrated solid but lower performance with an accuracy of (77.72%) and an F1-score of (0.7717), while CRNN showed a noticeable decline in performance with (59.39%) accuracy and a lower F1-score of (0.5874), suggesting difficulties in generalizing. CRNN with Attention performed the worst, with an accuracy of (50.61%), precision of (0.5162), recall of (0.5060), and an F1-score of (0.4951), indicating that the attention mechanism did not significantly enhance the model's ability to classify accurately. Overall, the results highlight VGG19 as the most effective model for the classification task, followed by VGG16, while the CRNN-based models, especially CRNN with Attention, struggled to match their performance. The comprehensive evaluation across all metrics suggests that VGG-based architectures are better suited for this task, providing superior accuracy and balance in predictions.

V. Conclusion and Future Work

Incorporating different data sources like audio-visual elements and lyrics could enhance the accuracy of music genre classification in future research. Advanced neural network structures, such as merging CNNs, CRNNs, and attention mechanisms, could result in improved performance. Moreover, enhancing model interpretability by utilizing methods such as attention visualization and feature attribution analysis will reveal the thought process behind the decisions made by these models. Options such as transfer learning and self-supervised learning may also tackle difficulties associated with less represented genres.[15]

Another crucial area to explore is the use of multimodal learning techniques, which merge various forms of data to develop stronger models. Improving instantaneous categorization for live-streaming platforms and creating fast methods for edge devices could boost the scalability and user-friendliness of genre classification systems. To sum up, progress in neural networks, data resources, and model interpretability will bring about substantial enhancements in music genre categorization, leading to more precise and customized music platforms.

References

[1] Yu-Huei Cheng, Pang-Ching Chang, Duc-Man Nguyen, and Che-Nan Kuo. "Automatic Music Genre Classification Based on CRNN." In: *Engineering Letters* 29.1 (2020).

[2] Albert Jimenez and Ferran Jose. "Music genre recognition with deep neural networks". In: *Universitat Politècnica de Catalunya* (2018).

[3] Weibin Zhang, Wenkang Lei, Xiangmin Xu, and Xiaofeng Xing. "Improved music genre classification with convolutional neural networks." In: *Interspeech*. 2016, pp. 3304–3308.

[4] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. "Convolutional recurrent neural networks for music classification". In: *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 2392–2396.

[5] Keunwoo Choi, George Fazekas, and Mark Sandler. “Explaining deep convolutional neural networks on music classification”. In: *arXiv preprint arXiv:1607.02444* (2016).

[6] Ndiatenda Ndou, Ritesh Ajoodha, and Ashwini Jadhav. “Music genre classification: A review of deep-learning and traditional machine-learning approaches”. In: *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE. 2021, pp. 1–6.

[7] Hareesh Bahuleyan. “Music genre classification using machine learning techniques”. In: *arXiv preprint arXiv:1804.01149* (2018).

[8] Keunwoo Choi, George Fazekas, Mark Sandler, and Jeonghee Kim. “Auralisation of deep convolutional neural networks: Listening to learned features”. In: *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR*. 2015, pp. 26–30.

[9] Ahmet Elbir and Nizamettin Aydin. “Music genre classification and music recommendation by using deep learning”. In: *Electronics Letters* 56.12 (2020), pp. 627–629.

[10] Yigang Meng. “Music Genre Classification: A Comparative Analysis of CNN and XGBoost Approaches with Mel-frequency cepstral coefficients and Mel Spectrograms”. In: *arXiv preprint arXiv:2401.04737* (2024).

[11] Keunwoo Choi, George Fazekas, and Mark Sandler. “Automatic tagging using deep convolutional neural networks”. In: *arXiv preprint arXiv:1606.00298* (2016).

[12] Yang Ding, Hongzheng Zhang, Wanmacairang Huang, Xiaoxiong Zhou, and Zhihan Shi. “Efficient Music Genre Recognition Using ECAS-CNN: A Novel Channel-Aware Neural Network Architecture”. In: *Sensors* 24.21 (2024), p. 7021.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[14] Sergey Ioffe. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015).

[15] Juliano Henrique Foleiss and Tiago Fernandes Tavares. “Random Projections of Mel-Spectrograms as Low-Level Features for Automatic Music Genre Classification”. In: *arXiv preprint arXiv:1911.04660* (2019).