DOI: https://doi.org/10.4038/jrm.v3i2.49

Sinhala Music Genre Classification Using a Deep Learning Approach

Anusha Jayasiri[1*], L. D. Peellawalage[2]

JRM JOURNAL OF RESEARCH IN MUSIC
VOLUME 03 | ISSUE 02 | 2025

Abstract

Traditional approaches to genre classification have relied on user ratings, emotions, and genres, often utilizing Support Vector Machines (SVM) without leveraging deep learning methodologies for genre classification. To address this gap, this paper presents a comprehensive and novel investigation into the application of deep learning techniques based on Convolutional Neural Networks (CNNs) for Sinhala music genre classification using spectrogram images derived from Mel-frequency Cepstral Coefficients (MFCCs). The methodology involves data collection and preprocessing, model architecture design, training procedures, experimentation, and the presentation of results. A dataset comprising 325 Sinhala songs, with 65 songs representing each of five prominent genres (pop, hip-hop, baila, calypso, and classical), was compiled. From these audio files, 1300 MFCC spectrogram images were extracted for feature representation. Preprocessing steps included format conversion, audio segmentation, feature extraction, visualization, and normalization to prepare the data for input to the CNN model. The CNN architecture designed for genre classification consists of multiple convolutional and pooling layers followed by fully connected layers. Transfer learning with pre-trained models, including VGG-19, InceptionV3, EfficientNet, and DenseNet was explored. CNN and RNN hybrid models were explored by adding an LSTM layer followed by CNN layers. Hyperparameter tuning techniques, particularly Random Search, were employed to optimize model performance. Evaluation metrics such as accuracy, precision, recall, and F1-score were utilized to assess the effectiveness of the classification models. Results indicate promising performance, with the DenseNet model achieving the highest accuracy of 78.0%. However, challenges such as dataset limitations remain areas for improvement.

Keywords

Audio Segmentation, Convolution Neural Network, Sinhala Songs

[1] Department of Information Technology, University of Visual and Performing Arts, Sri Lanka. Email: anusha.j@vpa.ac.lk
[2]Department of Statistics and Computer Science, University of Kelaniya, Sri Lanka.

Introduction

Music has been an integral part of human culture, acting as a universal language that transcends cultural and geographical barriers. The classification of music into genres has long been a subject of interest in the fields of signal processing, musicology, and, more recently, machine learning. Understanding and categorizing music genres is not merely an academic pursuit, as it holds practical implications for content organization, recommendation systems, and the preservation of cultural heritage. Early methodologies focused on handcrafted feature extraction, capturing characteristics such as tempo, pitch, and spectral content to identify musical genres. As technology advanced, machine learning techniques emerged, allowing algorithms to learn discriminative features directly from the data. In recent years, deep learning has emerged as a powerful tool, and specifically, Convolutional Neural Networks (CNNs) have demonstrated remarkable success in various image and audio-related tasks, making them an attractive choice for music genre classification. Despite the progress in this field, a significant research gap remains when it comes to the classification of the Sinhala music genres. From traditional folk music rooted in rural communities to contemporary fusion genres influenced by global musical trends, Sinhala music reflects the dynamic interplay between local traditions and external influences. Sinhala music mostly falls into categories such as folk music (kavi/virindu), Hindustani ragadari music, baila, calypso, classical, and western (pop, hip-hop) music. In contemporary Sinhala music, we can primarily identify it under five categories: pop, hip-hop, baila, calypso, and classical.

Sinhala music resulting from diverse cultural influences has become challenging to classify into different genres. This issue affects recommendation engines, content organization, and cultural preservation. Manual classification is difficult due to the complex nature of Sinhala music. The complexity inherent in Sinhala music makes manual classification a challenging task, hindering listeners from finding music they enjoy. Existing models, trained in Western music, struggle with Sinhala genres, leading to misclassification. Consequently, there is no standard framework for categorizing Sinhala music genres.

This research aims to develop a deep learning model for Sinhala music genre classification, capable of capturing the unique characteristics of Sinhala music genres and to develop a CNN-based model for classifying Sinhala music into 5 prominent genres: baila, calypso, classic, pop,

and hip-hop. It developed a sophisticated deep-learning model tailored for classifying Sinhala music genres by leveraging a comprehensive dataset, advanced feature extraction techniques, and cutting-edge model architectures to achieve robust performance in classifying Sinhala music into prominent genres.

The significance of this research lies in its multifaceted contribution to the domain of Sinhala music genre classification. Through meticulous dataset compilation and the advancement of classification techniques, this study addressed critical gaps in existing methodologies. Delving into the preservation and recognition of Sinhala music, it not only enriches the understanding of this vibrant musical tradition but also ensures its longevity in the digital landscape. Furthermore, the research addressed the absence of a proper dataset for classifying Sinhala songs by compiling a comprehensive dataset covering specified genres. This dataset serves as a crucial resource for future studies and applications in Sinhala music genre classification. The research contributes to the preservation and recognition of Sinhala music by providing a means to automatically categorize and organize it. By accurately classifying Sinhala music genres, the research facilitates the discovery, exploration, and appreciation of this rich musical tradition, ensuring its continued relevance and resonance in the digital age. The development and application of a deep learning model in a culturally specific context highlights the adaptability of machine learning techniques to diverse cultural domains. By showcasing the efficacy of deep learning models in accurately classifying Sinhala music genres, the study underscores the potential of machine learning in understanding and preserving cultural heritage, fostering cross-cultural dialogue, and promoting cultural diversity in the field of music genre classification.

Literature Review

This literature review synthesizes findings from several studies focusing on different aspects of music classification, particularly in the context of Sinhala songs.

Peiris and Jayarathne (2016) explored musical genre classification in Sri Lankan music using audio signals. They employed two feature vectors and a Support Vector Machine Classifier with a radial-basis kernel function to automatically classify music genres. Their approach captured various audio features, including frequency, temporal, cepstral, and modulation frequency domains, achieving the highest accuracy of 74.5% on a dataset of 108 songs across the five genres: Baila, Classical, Ragadari, and Calypso. The audio processing involved

dividing the signal into frames, applying the Windows function, FFT, Mel-Filters, and MFCC extraction. They addressed genre classification challenges using the One vs One(OVO) strategy and experimented with tree-based classifiers, SVM with PCA, MLP, and J48 decision trees. Boosting and bagging techniques were also explored, concluding boosting that is more effective with weak classifiers and less noisy data (Peiris and Jayaratne, 2016).

Dhanapala and Samarasinghe (2024) comprehensively reviewed studies on Music Emotion Classification (MEC) over the period from 2006 to 2023, focusing exclusively on research that utilized audio data. From an initial pool of 42 publications identified via Google Scholar, 20 papers were selected for detailed analysis. The authors categorized MEC methodologies across four dimensions: acoustic feature extraction, emotion modeling, classification techniques, and performance evaluation. Commonly used acoustic features included rhythm, pitch, timbre, spectral characteristics, and harmony, while prevailing emotion categories were happiness, anger, sadness, and relaxation. Support Vector Machines (SVMs) emerged as the most frequently used machine learning algorithm, although studies also experimented with regression, neural networks, and fuzzy logic approaches (Dhanapala & Samarasinghe, 2024).

Abeyratne and Jayarathna (2019) conducted a study on the emotion-based classification of Sinhala songs using Music Emotion Recognition, a subset of Music Information Retrieval. They experimented with 18 different feature selection algorithms and supervised machine learning techniques. The study used a dataset of 162 labeled Sinhala songs, categorized into six emotion categories (excitement, happiness, sadness, sleepiness, relaxed, and calmness) using Russel's 2D Valence-Arousal Model. Audio features were extracted from the first 90 seconds of each song, and three feature selection methods (Relief, Information Gain, and Correlation-based Feature Selection) were evaluated. The Random Forest Algorithm, paired with the Relief Feature Selection Method, achieved the highest accuracy of 91.98% in classifying emotions. The study concluded that this combination was the most effective for classifying emotions in Sinhala songs, with all tested machine learning algorithms achieving over 85% accuracy. (Abeyratne and Jayaratne, 2019)

Lakshitha and Jayaratne (2016) investigated the role of melody in predicting emotions in Sinhala songs. Despite initial low classification accuracy, they experimented with feature selection and an ensemble classification technique to improve results. Using a dataset of 158 songs labeled with emotions like happiness, excitement, sadness, calmness, and heroic feeling, they initially achieved the best accuracy of only 45.57% with Naïve Bayes, particularly

struggling with distinguishing emotions. Feature selection and ensemble methods, such as bagging and boosting, were also tested, with the highest accuracy reaching 46.2% using a combination of Naïve Bayes and LibSVM. They then incorporated additional features related to rhythm and timber, which improved performance for most algorithms, except LiSVM (Lakshitha and Jayaratne, 2016)

Costa et al (2011) and Nirmal et al (2020) have introduced a novel approach to music genre classification by converting audio signals into spectrograms and treating them as texture images. The method involves time decomposition to create sub-signals, which are then transformed into spectrograms representing different parts of music pieces (start, middle, and end). Features are extracted using Gray Level Co-occurrence Matrix (GLCM) descriptors, and a Zoning mechanism captures non-uniform texture, resulting in multiple feature vectors for each spectrogram. The Support Vector Machine Classifier, with normalized and optimized features, is used for classification, employing 3-fold cross-validation and a majority voting scheme across 10 zones, which provided the best recognition rate. Experiments conducted on a subset of the Latin Music Database showed a 7% improvement in recognition accuracy when this texture-based classifier was combined with others trained on short-term, low-level audio characteristics, using the Max rule. (Nirmal et al., 2020); (Costa et al., 2011)

Choudhury et al (2023) proposed a CNN-based algorithm for music genre classification that uses spectrograms generated by Short-time Fast Fourier Transform (STFT) to capture dynamic music components. This approach improves upon traditional methods like MFCC by creating dynamic spectrograms that better represent music's temporal features. A modified CNN, with specialized filters for identifying components like percussion and harmonic elements, is used to extract detailed features. Convolution operations create feature maps, which are then subsampled to reduce dimensionality and improve robustness to pitch and tempo variations. These high-level features are fed into a multi-layer perceptron (MLP) classifier. The algorithm was tested on the Tzanetakis1 dataset, consisting of 1,000 audio clips across 10 genres. Various feature extraction methods were compared, with the CNN-based COV_FFT method achieving the highest accuracy of 72.4%. (Choudhury et al., 2023)

Dong proposed a method for music genre classification using a Convolutional Neural Network (CNN) inspired by human perception and auditory system neurophysiology, involving training a CNN on short music segments and combining predictions for genre classification. The CNN

model uses Mel-spectrograms as input, achieves human-level (70%) accuracy, and the learned filters resemble septotemporal receptive fields in the auditory system. (Dong, 2018)
Vishnupriya and Meenakshi (2018) developed an approach for automatic music genre classification using Convolutional Neural Networks to efficiently index and organize the growing number of music tracks. Their method involves preprocessing songs to extract Mel Frequency Cepstral Coefficients as feature vectors, which are then entered into a CNN for classification. They tested their approach on a dataset with ten genres, using 800 songs for training and 200 for testing. The results showed promising accuracy levels of 76% for Mel Spectrogram features and 47% for MFCC features.

Ahmed et al (2020) proposed the use of Convolutional Neural Networks (CNNs) for classifying musical genres using the Marsyas Audio Dataset. It emphasizes the importance of music genre classification and highlights the limitations of traditional methods compared to CNNs. They reviewed previous research efforts in sound classification and noted the absence of CNN applications specifically for the Marsyas Audio dataset. The model achieves a classification accuracy of 92%, demonstrating the effectiveness of CNN with MFCC spectrograms for genre classification.

Xie et al (2012) developed an MFCC_E algorithm that enhances the Mel Frequency Cepstrum Coefficient Method for abnormal audio recognition. While MFCC is effective, it is computationally demanding. limiting its hardware implementation. MFCC_E reduces computation by 50% compared to MFCC while maintaining similar recognition rates. The process includes preprocessing, feature extraction, and classification using models such as SVM, HMM, or GMM.MFCC_E which simplifies computation through adjustments in the framing modules and a modified pre-emphasis parameter. Experimental results show MFCC_E performances comparable to MFCC, making it more suitable for hardware use.

Huang et al (2017) introduced DenseNet, a groundbreaking convolutional network architecture that connects all layers directly in a feed-forward manner. This design enhances feature propagation, promotes feature reuse, and reduces parameter redundancy. DenseNet outperforms state-of-the-art models on benchmark datasets like CIFAR-10, CIFAR-100, SVHN, and ImageNet, while using fewer computational resources. It achieves accuracy comparable to ResNet but with significantly fewer parameters, addressing scalability and efficiency issues. The paper also explores the theoretical aspects of DenseNet, highlighting its

compactness, deep supervision, and connection strategies. DensNet marks a significant advancement in convolutional network design.

Methodology

Earlier attempts to classify Sinhala songs were based on user ratings, emotions, and genres. SVM was primarily used for genre classification without employing the use of deep learning techniques. The study involves the application of machine learning techniques, particularly Convolutional Neural Networks, for the classification of Sinhala music genres based on spectrogram images derived from Mel-frequency cepstral coefficients (MFCCs).

When processing audio signals, Mel-frequency cepstral coefficients (MFCCs) are an often-utilized characteristic. They are especially good at capturing music content since they capture the spectral characteristics (frequency content and temporal variations) of audio signals. MFCCs are particularly useful in tasks like speech recognition, speaker identification, and music genre classification.
The following sections briefly explain the other methodology parts:

Description of the Dataset:

To facilitate the development and training of the Convolution Neural Network(CNN) model for Sinhala music genre classification, a comprehensive dataset of Sinhala songs was compiled. The dataset consisted of 325 songs, with 65 songs representing each of the five prominent genres: pop, hip-hop, baila, calypso, and classical. The audio files were sourced from diverse sources, including online music platforms, digital archives, and local music collections. The dataset was collected from various sources, including online music databases, music streaming platforms, and personal collections of Sinhala music. Careful consideration was given to ensure diversity in genre representation and audio quality across the dataset.

The dataset used in this research consists of 1300 MFCC (Mel-frequency cepstral coefficients) spectrogram images extracted from Sinhala music audio files. Each audio file was represented as a sequence of frames, and MFCC spectrograms were computed for each frame to capture its frequency characteristics.

Data Preprocessing:

Before extracting features from the audio data, the songs were preprocessed to ensure consistency and compatibility with the CNN model. The first step in data preprocessing

involved converting the audio files into a suitable format for analysis. The MP3 audio files were converted to WAV format using the PyDub library, and 30-second segments were extracted from each song to standardize the duration for analysis. Feature extraction played a crucial role in capturing the distinguishing characteristics of Sinhala music genres. Mel-frequency cepstral coefficients (MFCCs) were extracted from the audio segments using the librosa library in Python. MFCCs represent the spectral features of audio signals and are commonly used in music genre classification tasks. A total of 15 MFCC coefficients were computed for each audio segment, capturing information about the timbral and spectral characteristics of the music.

Using the PyDub Python library, initial MP3 audio files were converted to WAV format to ensure compatibility with the feature extraction process. For audio segmentation, each audio file was segmented into smaller, fixed-length segments to ensure uniformity in the input data fed into the CNN model. For feature extraction, the primary feature extracted from the audio data was the Mel-frequency cepstral coefficient. In this case, 15 Mel-frequency cepstral coefficients were extracted from each audio segment.

For visualization, Spectrogram images were generated from the extracted MFCCs to provide a visual representation of the audio features. The spectrograms were represented as 2D images, with time on the horizontal axis and frequency on the vertical axis. These spectrogram images serve as the input data for the CNN model. For Normalization, the MFCC spectrograms were normalized to a fixed range, typically between 0 and 1, to ensure uniformity in feature scaling across different segments and audio files. After the data preprocessing stage, the dataset had 1300 MFCC spectrograms fed into the CNN model.

Model Architecture:

Model architecture plays a pivotal role in the success of Sinhala music genre classification. The CNN architecture is designed to effectively learn discriminative features from MFCC spectrogram images extracted from Sinhala music audio segments. The model optimization involves configuring hyperparameters such as learning rate, batch size, and optimization algorithms. techniques like Dropout Regularization and Batch Normalization are applied to mitigate overfitting and improve training convergence.

The CNN architecture consisted of convolutional layers with varying filter sizes to capture local and global features of the input spectrograms. Max-pooling layers were employed to

downsample the feature maps and reduce computational complexity while preserving important features. ReLU activation functions were used to introduce non-linearity to the model and learn complex patterns from input data. The final dense layers comprise a SoftMax activation function to output probability distributions over the target genres. The model was trained using the categorical cross-entropy loss function and optimized using the Adam optimizer with a learning rate of 0.0001. Additionally, applied callbacks, early stopping, and learning rate scheduling.

Transfer Learning with Pre-Trained Models:

In addition to using custom CNN architecture, several transfer learning models were explored to leverage pre-trained convolutional neural networks for genre classification. Transfer learning is a machine learning technique where a model trained on one task is repurposed for another related task, and it can accelerate the learning process for the target task, especially when the target task has limited data. For this Sinhala music genre classification, transfer learning can be applied using pre-trained convolutional neural network (CNN) models that have been trained on large-scale image datasets such as ImageNet. DenseNet121, EfficientNet, InceptionV3, and VGG19 pre-trained models which were also used for this study.

CNN-LSTM Hybrid Model:

The CNN-LSTM model architecture extends the CNN model by adding a Long Short-Term Memory (LSTM) layer with 128 units after the convolutional layers to capture temporal dependencies in the input spectrogram sequences. This LSTM layer has been added between the flatten layer and the dense layer as the reshape layer and the LSTM Layer.

Training Procedure:

The training process involves optimizing the CNN model and Transfer Learning model parameters using the backpropagation algorithm and an optimization algorithm such as the Adam Optimizer. The model was compiled with an appropriate loss function and with evaluation metrics to measure performance during training. Hyperparameters were tuned using Random Search to find the optimal configuration that maximizes model performance. The training data was passed through the network for multiple epochs, and each epoch consisted of one pass through the entire dataset. The validation dataset was used to monitor the model's performance and detect overfitting. Early stopping and learning rate scheduling criteria were employed to halt training if the validation loss fails to decrease for a certain number of epochs.

Results and Discussion

The research implementation environment was conducted using Google Colab, a cloud-based platform that provides free access to GPU resources for running machine learning experiments, and Python language used to implement the model and use relevant Python libraries.
The dataset-splitting process was performed within the Google Colab environment. 1300 MFCC images were divided into training, validation, and testing sets. These splits were randomly generated using a random library.

Evaluation metrics are essential for assessing the performance of the developed model accurately. For this research, several evaluation metrics such as Accuracy, Precision, Recall, and F1-Score were employed to measure the effectiveness and efficiency of the classification model.

The results section discusses the performance of various models, including Convolutional Neural Networks (CNNs), transfer learning models such as VGG19, InceptionV3, EfficientNet, and DenseNet, as well as hybrid models, combining these with Long Short-Term Memory (LSTM) networks to leverage their strengths in processing sequential data. The study compares these models in classification tasks using evaluation metrics such as accuracy, precision, recall, and F1-score.

CNN Model:

The CNN model's performance in classifying the Hip-hop classification achieved high precision (0.79) and recall (0.85), resulting in an F1-score of 0.81. Other genres like 'baila', 'calypso', and 'classical' demonstrated reasonable performance with precision and recall scores ranging from 0.75 to 0.82. However, the model struggled with the 'pop' genre, with a precision of 0.53 and a recall of 0.69, resulting in an F1-score of 0.60. Overall accuracy was 0.72, indicating the correct classification of around 72% of samples.

CNN Model with LSTM:

The hybrid CNN+LSTM model exhibits significant improvements in precision, recall, and F1-score compared to a standalone CNN model, particularly excelling in 41 classifying 'classical' and 'hip-hop' genres. 'Classical' classification achieved a precision of 0.91, a recall of 0.77, and an impressive F1-score of 0.83. Similarly, the hip-hop classification demonstrates strong precision (0.80) and recall (0.92), resulting in an excellent F1-score of 0.86. However, some

genres like 'baila' and 'pop' show slightly lower performance compared to the CNN-only model. Overall, the hybrid model maintains an accuracy of 0.72, correctly classifying approximately 72% of samples across all genres.

VGG19 Model:

The VGG19 model demonstrates competitive performance in classifying Sinhala music genres, achieving an overall accuracy of 0.69. 'Classical' genre classification shows strong precision (0.77), recall (0.77), and F1-score (0.77), reflecting accurate identification. 'Hip-hop' classification stands out with high recall (0.92) and an F1-score of 0.77. However, the 'baila' and 'calypso' genres exhibit slightly lower performance, with precision, recall, and F1-score around 0.64. Similarly, the 'pop' genre classification shows moderate performance, with a precision of 0.70, a recall of 0.54, and an F1-score of 0.61.

VGG19 Model with LSTM:

The VGG19+LSTM hybrid model performs well in classifying Sinhala music genres, achieving an accuracy of 0.66. 'classical' consistently achieves 0.77 across these metrics, indicating accurate identification. 'Hip-hop' classification stands out with a recall of 0.85 and an F1-score of 0.73. However, the 'baila' and 'calypso' genres exhibit slightly lower performance, with scores around 0.59 and 0.58, respectively. 'Pop' genre classification shows a precision of 0.70 and a recall of 0.54, resulting in an F1-score of 0.61.

EfficientNet model:

The EfficientNet model achieved robust performance in classifying Sinhala music genres with an accuracy of 0.74. Its efficient architecture efficiently captures features from spectrogram images, leading to accurate genre classification. The model maintains balanced precision, recall, and F1 scores, ranging from 0.57 to 0.83.

EfficientNet model with LSTM:

The EfficientNet+LSTM model achieved competitive performance with an accuracy of 0.71. Notably, the 'classical' genre classification stands out with high precision, recall, and F1-score of 0.86, 0.92, and 0.89, respectively, showcasing the model's proficiency. The hip-hop genre also demonstrates notable performance, with scores ranging from 0.79 to 0.85. However, challenges exist in classifying genres like 'pop', where precision, recall, and F1-score are comparatively lower.

Inception V3 Model:

This model achieved an accuracy of 0.62 in classifying the Sinhala music genre; ''classical' music, with high precision, recall, and F1-score of 0.83, 0.77, and 0.80, respectively. The hip-hop genre also showed decent performance, with scores ranging from 0.67 to 0.77.

Inception V3 Model with LSTM:

Using this model achieved an accuracy of 0.65 in classifying Sinhala music genres. It showed strong performance in classifying 'classical' music, with precision, recall, and F1-score of 0.83, 0.77, and 0.80, respectively. 'Hip-hop' classification also performed decently, with scores ranging from 0.67 to 0.77.

DenseNet Model:

This model achieved an overall accuracy of 0.78. Across various genres, including 'baila', 'classical', and 'hip-hop', the model showed high precision, recall, and F1-scores ranging from 0.80 to 0.91, highlighting its robust classification capability.

Dense Net Model with LSTM:

In this case, achieved an overall accuracy of 0.74, including 'baila' and 'classical', the model maintains balanced precision, recall, and F1-scores, ranging from 0.73 to 0.83. Additionally, the 'calypso', 'hip-hop', and 'pop' genres also exhibit respectable performance, with scores ranging from 0.57 to 0.82.

The following

Table ***1*** shows the overall precision, recall, and accuracy of each model for Sinhala song classification.

Table 1:Comparison of Models

Model	Precision	Recall	F1-Score	Accuracy
CNN	0.74	0.72	0.73	0.72
CNN+ LSTM	0.73	0.72	0.72	0.72
VGG19	0.69	0.69	0.69	0.69
VGG19+ LSTM	0.66	0.66	0.66	0.66
DenseNet	0.79	0.78	0.78	0.78

DenseNet+ LSTM	0.75	0.74	0.74	0.74
EfficientNet	0.75	0.74	0.74	0.74
EfficientNet+ LSTM	0.71	0.71	0.70	0.71
InceptionV3	0.62	0.62	0.61	0.62
InceptionV3+ LSTM	0.65	0.65	0.64	0.65

One key observation from the experimentation phase is the varying performance of different model architectures. While the CNN model provided a baseline, more sophisticated architectures DenseNet, VGG19, InceptionV3, and EfficientNet, demonstrated higher accuracy levels, suggesting the importance of leveraging pre-trained models for feature extraction. Exploration of a CNN with LSTM captured sequential patterns in the MFCC spectrogram sequences and obtained an overall accuracy of 72%. Pre-trained models like VGG19 and DenseNet performed with an accuracy of 69% and 78% respectively. Models like EfficientNet and InceptionV3 achieved accuracies of 74% and 62% respectively.

Conclusion

This research explores the automation of Sinhala music genre classification using Convolution Neural Network transfer learning with trained models and hybrid architectures. CNN models demonstrated strong baseline accuracy, with the CNN-LSTM hybrid slightly improving performance by capturing sequential audio patterns. Pre-trained models like DenseNet and EfficientNet outperformed traditional CNNs while LSTM integration into these models showed mixed results, emphasizing the need for careful model selection. Despite promising results, challenges like a particularly limited dataset size remain. Future research should focus on hyperparameter tuning, data augmentation, and ensemble learning to enhance classification accuracy. Continued exploration and collaboration with experts are recommended to develop robust and accurate systems for Sinhala music genre classification.

References

Abeyratne K. M. H. B., Jayaratne K. L. (2019). Classification of Sinhala Songs based on Emotions. 19th International Conference on Advances in ICT for Emerging Regions (ICTer), (pp. 1–10). Colombo, Sri Lanka.

Ahmed M.S., Mahmud M.Z., Akhter S. (2020). Musical Genre Classification on the Marsyas Audio Data Using Convolution NN. 23rd International Conference on Computer and Information Technology (ICCIT). DHAKA, Bangladesh: IEEE.

Choudhury N., Deka D., Sarmah S., Sarma P. (2023). Music Genre Classification Using Convolutional Neural Network. 4th International Conference on Computing and Communication Systems (I3CS). Shillong, India: IEEE.

Costa Y. M. G., Oliveira L. S., Koericb A. L. , Gouyon F. (2011). Music Genre Recognition Using Spectrograms. 18th International Conference on Systems, Signals and Image Processing. Sarajevo, Bosnia and Herzegovina: IEEE.

Dhanapala, S., Samarasinghe, K. (2024). Music emotion classification: A literature review. Journal of Research in Music, 2(2), 14–28.

Dong, M. (2018). Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification. Conference on Cognitive Computational Neuroscience. Philadelphia, Pennsylvania.

Huang G, Liu Z., Van Der Maaten L., Weinberger K. Q. (2017). Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition, (pp. 4700-4708). Honolulu.

Lakshitha M.G.V, Jayaratne K.L. (2016). Melody Analysis for Prediction of the Emotions Conveyed by Sinhala Songs," in 2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), Galle, Sri Lanka. IEEE International Conference on Information and Automation for Sustainability (ICIAfS). Galle, Sri Lanka.

Nirmal M R, Shajee Mohan B S. (2020). Music Genre Classification using Spectrograms. International Conference on Power, Instrumentation, Control and Computing (PICC). Thrissur, India: IEEE.

Peiris R., Jayaratne L. (September 2016). Musical Genre Classification of Recorded Songs based on Music Structure Similarity. European Journal of Computer Science and Information Technology, Vol 4, No.5, pp.70-88.

Vishnupriya S.; Meenakshi K. (2018). Automatic Music Genre Classification using Convolution Neural Network. International Conference on Computer Communication and Informatics (ICCCI). Coimbatore, India: IEEE.

Xie C., Cao X., He L. (2012). Algorithm of Abnormal Audio Recognition Based on Improved MFCC. Procedia Engineering, Vol. 29, pp. 731–737.