IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Harmonising Genres by Exploring Music Genre Classification through Comparative ML Models

Dr V Rajalakshmi[1], Dr T Padmavathy[2], Ms G R Khanaghavalle[3] , Ms S Keerthana[4]

[1,3] Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Tamil Nadu.

[2]Department of Database Systems, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore.

[4] Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai

Doi: https://doi.org/10.56975/ijcrt.v14i7.308993

Abstract :

Music genre classification is a challenging task in the field of audio signal processing and machine learning. As the volume of digital music continues to grow exponentially, the need for automated systems to organise and categorise this vast array of music becomes crucial. This research presents a comprehensive Music Genre Classification System (MGCS) that leverages advanced machine learning techniques to accurately categorise music into predefined genres. The proposed MGCS employs a feature extraction process to capture key characteristics of audio signals, including spectral features, rhythm patterns, and temporal information. These features serve as input to a well-designed machine learning model, such as a convolutional neural network (CNN) which is trained on a diverse and extensive dataset of labelled music samples. The evaluation of MGCS involves testing its performance on various benchmark datasets, comparing its accuracy, precision, recall, and F1 score with existing state-of-the-art methods. The results demonstrate the efficacy of the proposed system in achieving high classification accuracy across a wide range of music genres. Furthermore, the research explores the interpretability of the model's predictions, shedding light on the features that contribute most to genre classification decisions.

1. Introduction

Music genre labels are valuable classification schemes for arranging and grouping tracks, albums, and musicians into more inclusive groupings based on shared musical attributes. Genres are commonly used in music classification, both in physical and streaming music stores. Thus, the topic of automatically classifying music genres has received a lot of attention. Users are used to browsing through music in genre categories whenever they access the Internet because music has been categorised by genre for a long time. According to research by Lee and Downie [2004], target end users are more inclined to browse by genre than by recommendation, artist similarity, or even musical similarity.

The music signals are represented as pulse-coded digital signals modulated and stored as discrete-time signals. These audio signals are stored in their original form or a compressed form. The most fundamental primary features that define the music are dynamics, form, harmony,melody, pitch, rhythm, tone,timbre and tempo. The secondary features defining music including aesthetics, style, structure, and texture are formed on the permutation of various primary features.

Music genre classification refers to the process of classifying music into genres based on the above-mentioned primary and secondary features. Thus features that compose the music play a vital role in the classification system. The extraction of these features from the music falls under any one of the following categories: content-based extraction, symbolic extraction, text-based extraction, and reference-based extraction.

Large e-commerce sites, like Amazon, employ genre to group material in their catalogues so that customers may search for products according to their favourite genre. When these catalogues are added to the database, they are typically manually labelled. The automatic classification system will mislabel music if the media samples are not properly labelled. Due to its underlying truth, other similarity-based approaches—such as mood or composer-based similarity—will become as obscure between classes as genre classification.

A good music genre classification system thus poses two requirements: 1. Feature selection and 2. Selection of the classification model. The first section, or obtaining an appropriate feature, is frequently the underlying classification system's performance-controlling element (Yu et al., 2020). To categorise music content, researchers have also experimented with a variety of musical and non-musical elements. To achieve state-of- the-art (SOTA) results on several widely recognized datasets, many researchers have integrated these characteristics with conventional machine learning classifiers like KNN (Cover & Hart, 1967), GMM (Duda, Hart, et al., 2006), SVM (Boser, Guyon, & Vapnik, 1992), etc. To obtain a set of robust feature(s), a generalised Music Genre Classification system is indeed needed.

This paper is organised as follows: Section 2 deals with the related work in the classific of music genre systems based on the features used for classification. Section 3 deals with the basics behind the research work. The proposed system model and the results of the proposed model are discussed in Section 4.

2. Related Work

The application of various machine learning algorithms and approaches to increase the accuracy of music genre classification has been the subject of numerous studies in recent years. In this section, we'll examine some of the most important and pertinent studies on the classification of musical genres.There has been an interesting attempt towards Music genre classification in the recent years and most of the methods focus on the use of CNNs for classification.

Choi et al. (2016) employed CNN to classify music genres based on a dataset of 2000 audio samples. The effectiveness of employing CNN for feature extraction from raw audio data was demonstrated by the suggested model, which reached 65% accuracy. Most of the researchers focussed on the use of Convolution Neural Networks (CNN) for music classification. Latin American Music Database (LMD), Costa et al., (2017) applied CNN using spectrograms lead to a high recognition rate of 92 % . This was followed by a series of experiments using CNN and spectrograms were reported for the ISMIR 2004 database with an accuracy of 86.7 %.

In a study by Gao et al. (2018), a large-scale dataset of 1.2 million songs was utilised to classify music genres using a deep learning approach. CNN and LSTM (Long Short-Term Memory) were combined in the suggested model to handle feature extraction and classification. 89.6% accuracy was attained by the suggested model on the extensive dataset.

Han et al. (2018) employed the GTZAN Dataset and a CNN-SVM hybrid technique to classify music genres. Compared to the conventional ways of classifying music genres, the proposed model's accuracy of 85.9% was noticeably higher. Zhu et al. (2019) classified music genres using a transfer learning approach. The study suggested a model that employed SVM for classification and pre-trained CNNs for feature extraction. 84.2% accuracy was attained by the suggested model on the GTZAN Dataset. Table 1 lists out a detailed analysis of the various works in music genre classification.

Table 1: Related works in Music Genre Classification

Authors	Features	Classification Algorithm	Accuracy (%)
Lidy [5]	Loudness, amplitude modulation, rhythm histogram, statistical spectrum descriptor, onset features, and symbolic features.	Linear SVM	76.80
Panagakis [6]	Spectro Temporal Features	Non-Negative Tensor Factorization	78.20
		High Order Singular Value Decomposition	77.90
		Multilinear Principal Component Analysis	75.01
Panagakis [7]	Multiscale spectro-temporal modulation features	Non-Negative Tensor Factorization Multilinear	80.47
		High Order Singular Value Decomposition	80.95
		Principal Component Analysis	78.53
Tzanetakis [8]	Spectral centroid, spectral rolloff, spectral flux, first five MelFrequency cepstral coefficients, low-energy feature, zero-crossing rate, means and variances, beat	GMM	61.00
		KNN	60.00

	histogram		
Basili [4]	Intervals, instruments, instrument classes, time changes, note extension	Naïve Bayes (NB)	62.00
		Voting Feature Intervals (VFI)	56.00
		J48 - Quinlan algorithm	58.00
		Nearest neighbours (NN)	59.00
		RIPPER (JRip)	52.00
Bergstra [3]	Fast Fourier transform coefficients, real cepstral coefficients, Melfrequency cepstral coefficients, zero-crossing rate, spectral spread, spectral centroid, spectral rolloff, autoregression coefficients.	ADABOOST	82.34
Cataltepe [10]	Timbre, Rhythm and Beat Features	10-Nearest Neighbors	75.00
		LDA	86.33
Deshpande [11]	Spectrograms and Mel-frequency spectral coefficients	KNN	75.00
		SVM	90.00
Hamel [9]	Mel-scaled energy bands, octave-based spectral contrast features, spectral energy bands	Pooled Features Classifier	93.00
		Multi Time Scale Learning Model	94.30
Lambrou [12]	Mean, variance, skewness, kurtosis, accuracy, angular second moment, correlation, and Entropy	Least Squares Minimum Distance Classifier (Kurtosis vs. Entropy)	91.67
Xu [2]	Beat Spectrum, LPC derived cepstrum, zero-crossing rate, spectrum power, Mel-frequency cepstral coefficients	SVM	93.64

3. Background

3.1. Feature Analysis:

Features extracted from the raw audio can be classified into two categories namely time domain features and frequency domain features.

3.1.1. Time Domain Features

3.1.1.1. Amplitude Envelope (AE): The amplitude envelope is computed in the time domain. It refers to the maximum amplitude of all samples in a given timeframe T. It is a simple beat-related feature but very sensitive to outliers. It is defined as :

3.1.1.2. Root-Mean-Square-Energy: It is termed as RMS energy or RMS power which represents the new events in the audio segmentation. It is computed as

3.1.1.3. Zero - Crossing rate (ZCR): ZCR is the measure of the number of times the amplitude value changes from positive to negative within a time frame t. This feature is the most commonly used time domain feature in speech recognition. It acts like an indicator of pitch as higher ZCR represents higher frequency.

3.1.1.4. Tempo: Tempo is represented in terms of Beats Per Minute(BPM). It projects how fast or slow a music is; Different music will have different tempos. We aggregate the value by calculating the mean across all the frames.

3.2. Frequency Domain Features

3.2.1. Mel-frequency Cepstral Coefficient (MFCC):MFCC is a frequency domain feature derived from a signal's spectrogram. The cepstral of an audio signal forms the components of MFCC. The MFCC uses linearly spaced frequency bands which favours human auditory systems instead of using cepstral features alone. It captures the short-term spectral-based features. Mel-spectral vectors are decorrelated to generate MFCC by using discrete cosine transform.

3.2.2. Spectral Rolloff: Spectral Rolloff can be defined as the ratio of the distribution of the spectrum corresponding to the frequency value. It is used to distinguish harmonic and noisy sounds i.e, below roll-off frequency and above roll-off frequency. In general, the roll-off percent is 0.85.

3.2.3. Spectral Bandwidth :Spectral bandwidth gives the amplitude difference between brightness and frequency magnitude. It indicates the frequency range of the frame.

3.2.4. Spectral Contrast : Spectral contrast can be described as the decibel difference between the pit points and peaks in the spectrum. Spectral contrast enhancement plays an important role in the speech enhancement process. Over the past two decades, spectral contrast enhancement has gained popularity and huge improvement.

3.2.5. Spectral Centroid: Center of the gravity of the spectrum is known as the spectral centroid. It is inferred as the sharpness or brightness of the audio signal. The ratio between the spectral centroid and frequency is correlated with the brightness of the audio signal than the spectrum centroid alone.

3.2.6. Chroma : Chroma or Chromagram or pitch class profiles refer to the twelve different pitches in the music. Chromagram is an effective tool for analysing music. It

categorises the music into twelve pitch classes such as {C, C#, D, D#, E, F, F#, G, G#, A, A#, B}. The Chroma feature is used to extract the melodic and harmonic components of the music. The Chroma feature is very powerful in detecting the changes in timbre and instrumentation of music. A sequence of Chroma features is captured by taking the slicing window of the chroma feature for each frame. Chroma features can be calculated using many ways. Short-time Fourier transforms (STFT) with a combination of binning strategies are some choices to calculate the Chroma feature.

4. Proposed Methodology & Results

Each audio file is converted into mel-spectrogram using Librosa library which extracts mel-spectrogram with 128 mel-filter covering a frequency range between 0 to 22050 Hz. Mel-spectrogram is used as the input to the model. Mel-spectrogram is generated by applying a log scale to the frequency axis of Short Time Fourier Transform (STFT). The figure shows the generation of the mel-spectrogram from the audio. It is observed that the mel-spectrogram generated provides abundant textual information which will play a vital role in training a robust model for predicting the genre of the music. Mel-spectrogram of music signals are distinguishable for different genres.

K-Nearest Neighbour is a classification algorithm. Based on the training the given new points will be mapped to a class. First, the model finds the nearest neighbours in the feature space. Later this information is used to classify the new points to the respective classes. Using the local setting it approximates the point. This is the simplest form of classification since most of the classification is left until necessary classification. Though this is a simplest classification algorithm, it classifies the new point based on the closest neighbours. So most of the classification done by the model is fairly accurate. In MLP Classifier, the perceptron is considered as the fully connected layer with one output element. Perceptron was revived and feed forward neural networks were created by connecting two or more layers than a single layer perceptron. Then Multi Layer Perceptron was proposed which is a network of multiple perceptrons with an activation function. It also has a training algorithm named backpropagation. MLP is considered as a universal function approximator which can be used to construct classification and regression models. The fully connected layers are used to extract the features.

Support Vector Machines (SVMs) are a powerful class of supervised machine learning algorithms widely used in music genre classification systems. SVMs excel in identifying decision boundaries in high-dimensional feature spaces, making them well-suited for tasks where distinguishing between different genres is essential. In this context, audio features extracted from music, such as Mel-frequency cepstral coefficients (MFCCs) or chroma features, serve as input for the SVM model. By choosing an appropriate kernel, such as Linear or Radial Basis Function (RBF), SVMs effectively map input features to decision boundaries, enabling accurate classification of music genres. The interpretability of SVMs is valuable, as the decision boundaries can be visualised, aiding in understanding how the model discriminates between genres based on extracted musical attributes. Proper data preprocessing, model training, and evaluation contribute to creating an effective SVM-based music genre classification system, with opportunities for further exploration and optimization.

Convolutional Neural Networks (CNNs) have emerged as powerful tools for music genre classification, leveraging their ability to automatically extract hierarchical features from spectrograms, which represent the time-frequency distribution of audio signals. The network architecture comprises convolutional layers that convolve over the input spectrogram, capturing intricate patterns and features. Pooling layers subsequently reduce spatial dimensions, focusing on salient information. Fully connected layers then combine these features for high-level abstraction, leading to a final softmax layer that outputs predicted probabilities for different music genres. Training involves adjusting weights using labeled data, with techniques like data augmentation enhancing model robustness. CNNs excel at capturing both temporal and frequency dependencies in audio data, making them well-suited for discerning intricate patterns in music and

achieving high accuracy in genre classification tasks.

4.1 Experimental Setup and Results

The dataset used for our experiment is the GTZAN dataset provided by Tzanetakis and Cook [8]. This is one of the widely used benchmark datasets for music genre classification. It contains 1000 audio tracks that are evenly distributed into 10 different genres namely : Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, Rock. Each genre contains 100 audio tracks. Each audio track is sampled at a rate of 22.05 kHz at 16 bits as mentioned in table 2. The music files are 30s long with 16 bit PCM encoded in ‗.wav‘format. The dataset is approximately 1.2 GB in size. For our work, we splitted the dataset into 8/2 for training and testing purposes.

Table 2 : Description of the GTZAN dataset

Genre	Audio Tracks	Samples
Blues	100	66,179,400
Classical	100	66,211,610
Country	100	66,202,776
Disco	100	66,193,480
Hiphop	100	66,346,824
Jazz	100	66,223,328
Metal	100	66,159,680
Pop	100	66,150,400
Reggae	100	66,162,290
Rock	100	66,201,058

The wave plot of a single audio signal from each genre is pictorially represented in figure 1.

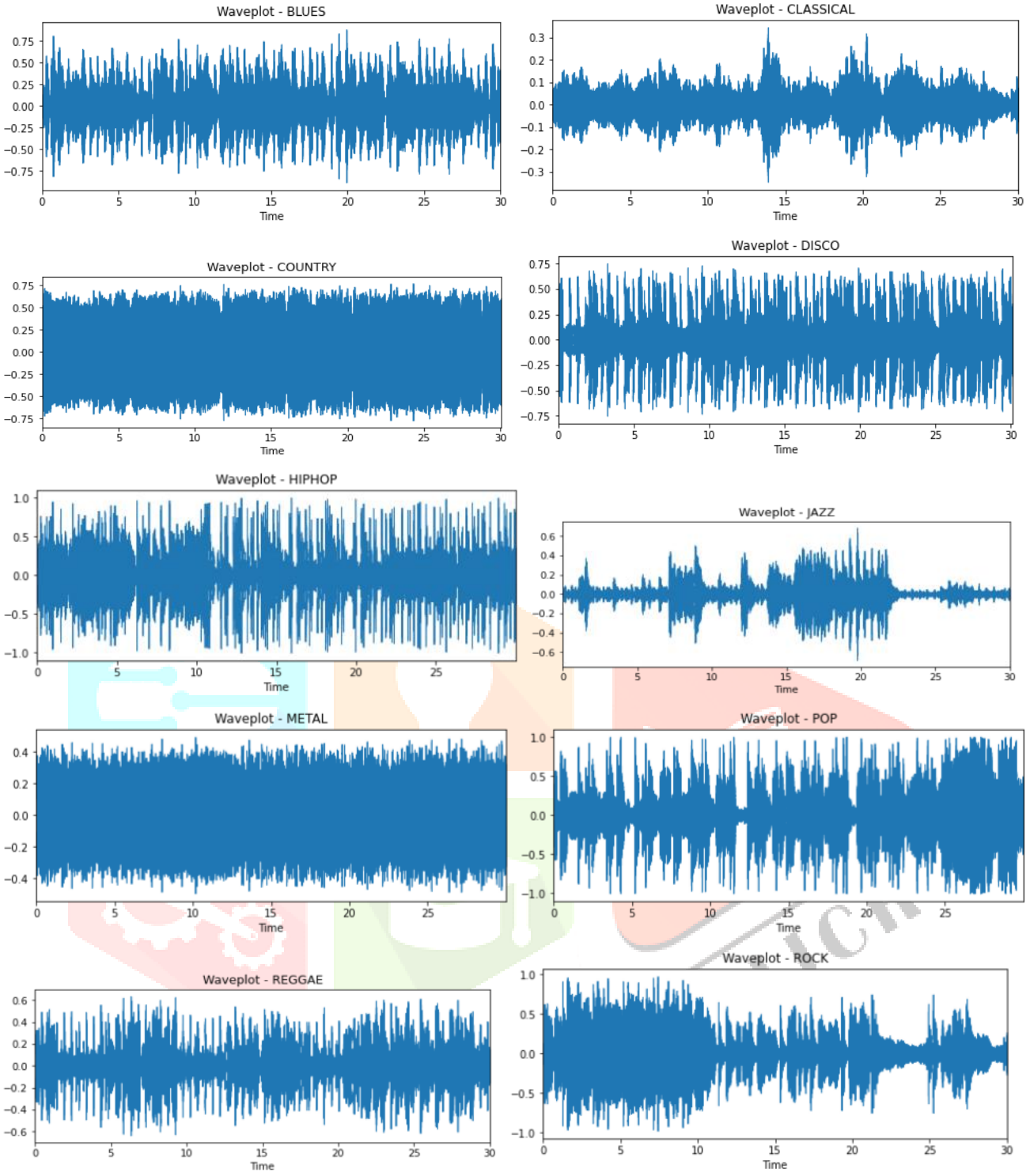
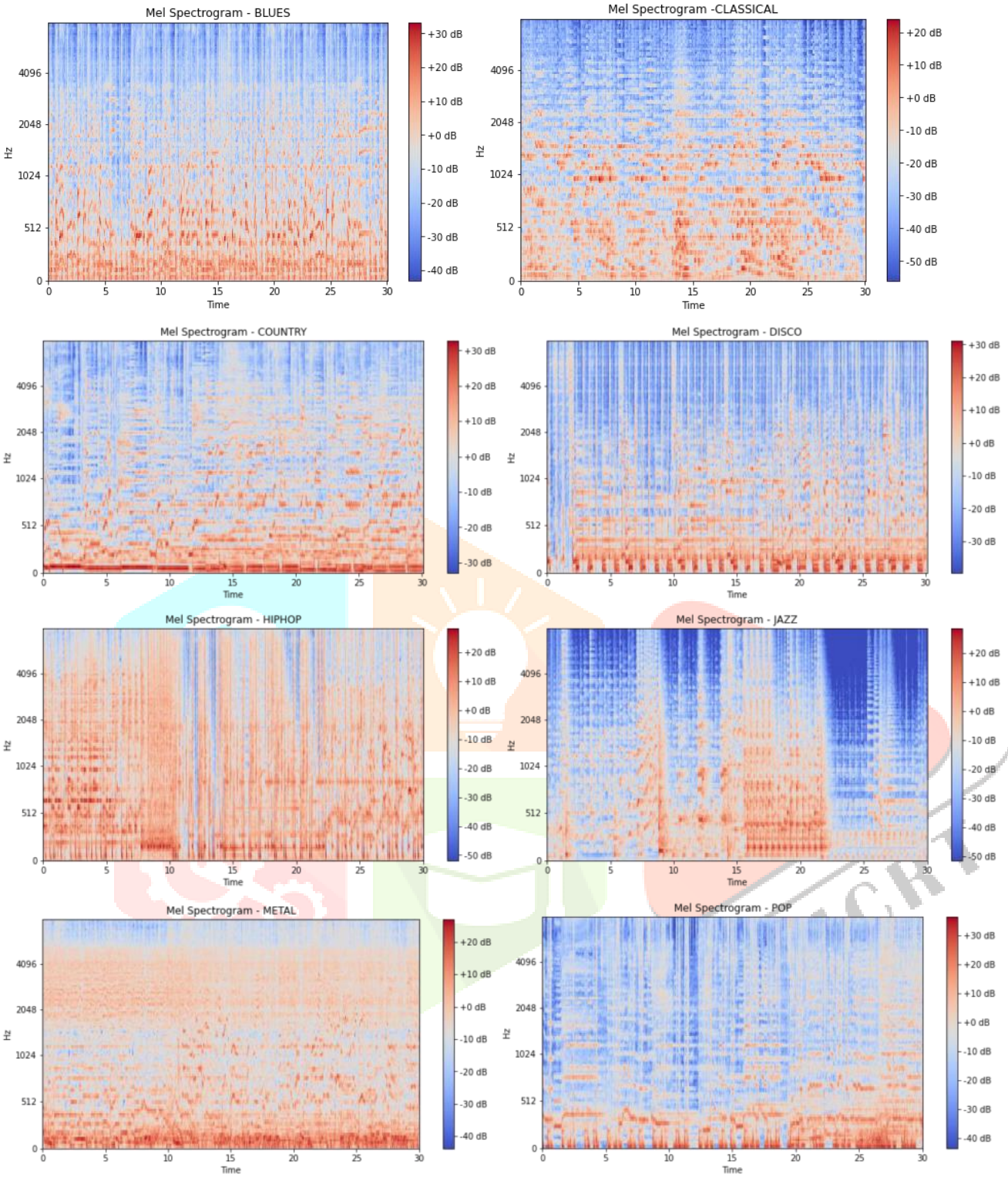


Figure 1 : Wave plot of a single audio signal from each genre


Mel Spectrogram - BLUES
Mel Spectrogram -CLASSICAL
Mel Spectrogram - COUNTRY
Mel Spectrogram - DISCO
Mel Spectrogram - HIPHOP
Mel Spectrogram - JAZZ
Mel Spectrogram - METAL
Mel Spectrogram - POP
Hz
Time

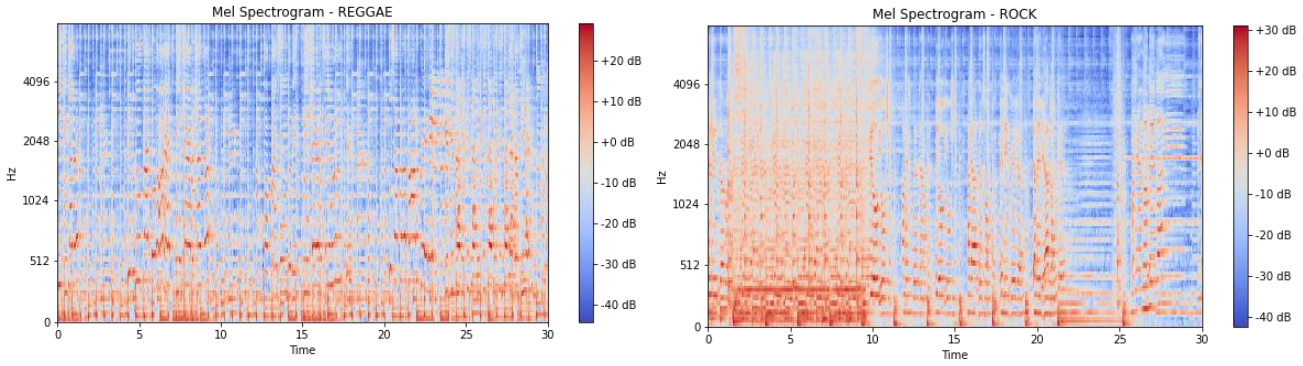


Figure 2 : Mel Spectrogram of a single audio signal from each genre

4.2 Evaluation metrics

To evaluate the performance of different algorithms we used the following metrics. They are Precision, Recall, F-measure and Accuracy scores. The upcoming sections explain this:

Precision: The Classifier‘s ability to return relevant instances is found using precision. It represents the quality of positive predictions made by the model. Precision is obtained by dividing the number of true positives by the total number of positive predictions.

Recall: Recall is also termed as Sensitivity. It measures the ability of the model to classify the positive samples. Recall is obtained by dividing the number of true positives by the total number of true positives and false negatives.

F- Measure: F-measure defines the quality of test accuracy of a model. It is calculated by combining precision and recall.

Accuracy: Accuracy is one of the popular metrics to evaluate the classification model. It is obtained by dividing the total number of correct predictions by the total number of inputs. When the dataset is balanced accuracy gives us a good indicator of our model.

4.3 Results and Discussion

In this investigation into music genre classification, four distinct models—K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN)—were employed to discern the most effective algorithm for accurate genre prediction. Among the

individual models, CNN demonstrated superior performance in terms of accuracy, consistently outperforming KNN, MLP, and SVM. This finding suggests that the hierarchical feature extraction capabilities of CNNs are particularly well-suited for capturing intricate patterns within the time-frequency domain of audio signals, crucial for accurate genre classification. The ablation study provided insights into the

contributions of each model. KNN served as a baseline, showcasing the importance of more sophisticated algorithms. MLP introduced non-linear transformations, SVMs brought versatility and interpretability, and CNNs excelled in automatically extracting hierarchical features from spectrograms. As a simple and intuitive algorithm, KNN offered a baseline level of accuracy. However, its performance may have been limited by the dataset's high dimensionality and the assumption of local similarities in the feature space. The non-linear transformations introduced by MLP contributed to capturing more complex relationships within the data. However, potential limitations might arise from the model's sensitivity to hyperparameter tuning and the risk of overfitting. SVMs, known for their versatility, provided a competitive accuracy. Their ability to handle non-linear decision boundaries and interpretability makes them valuable in understanding the discriminative features in music genre classification. The CNN demonstrated superior accuracy, showcasing the importance of convolutional operations in capturing intricate patterns within the time-frequency domain. This suggests that the model's hierarchical feature extraction is well-suited for discerning complex genre characteristics. The models were evaluated on a diverse dataset, and while CNN excelled in accuracy, it's important to consider factors such as generalisation to unseen data and robustness to variations in music styles and characteristics.

Table 3 : Performance accuracy of various models on GTZAN Dataset

Model	Accuracy
KNN	87%
MLP	88%
SVM	84%
CNN	**92%**

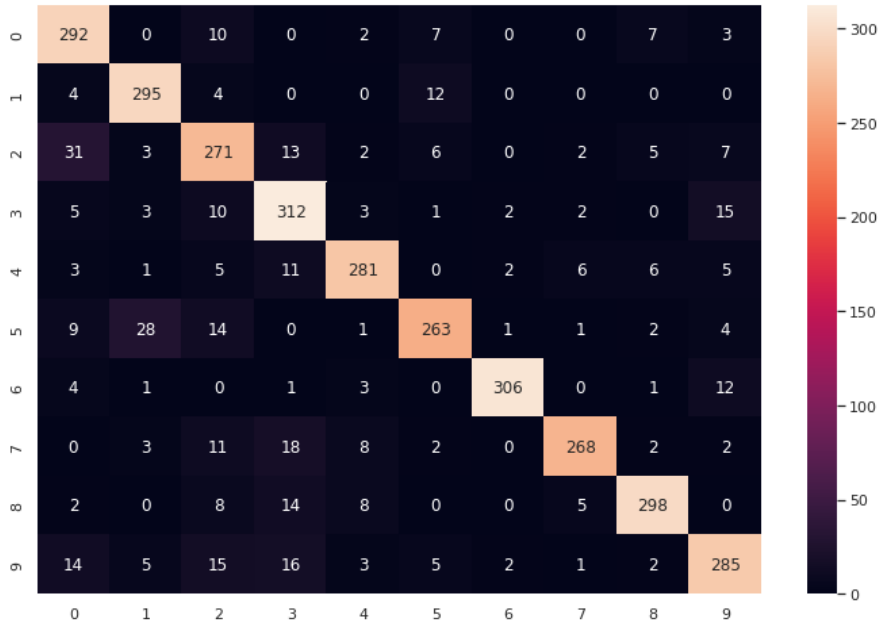


Figure 3 : Confusion matrix for KNN

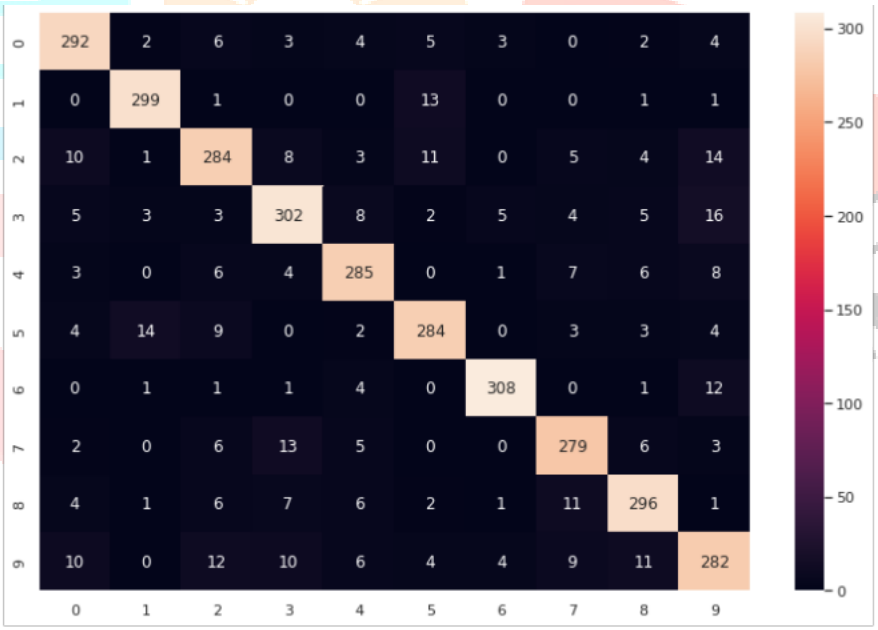


Figure 4 : Confusion matrix for MLP

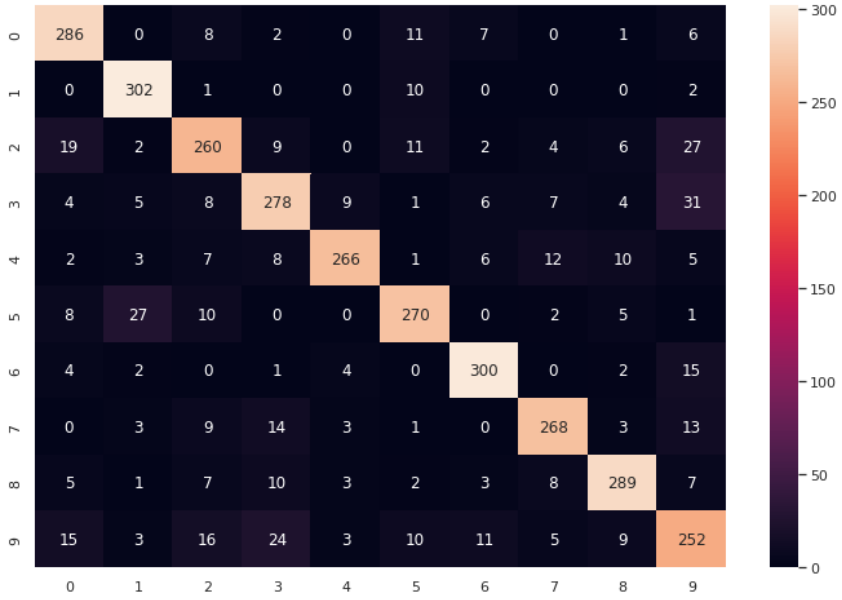


Figure 5 :Confusion matrix for SVM

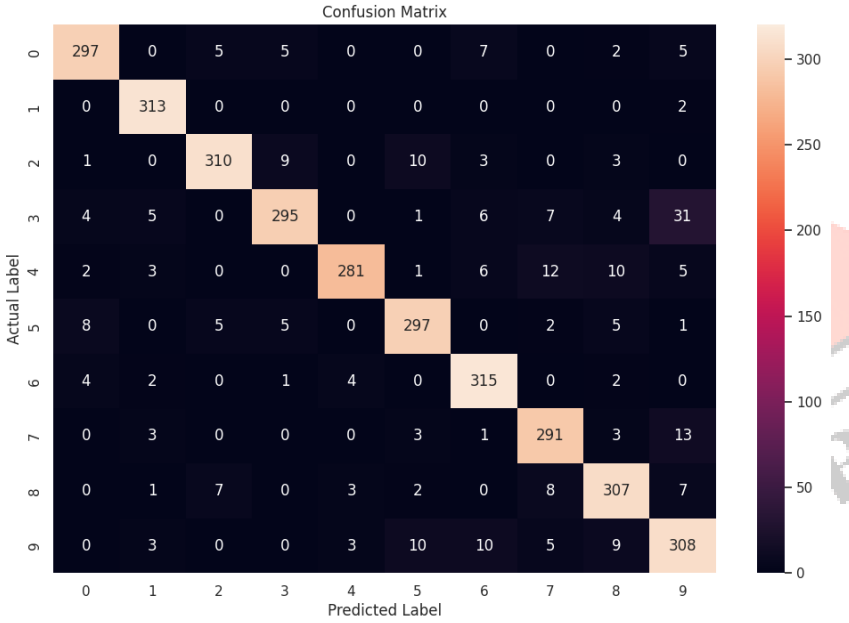


Figure 6 :Confusion matrix for CNN

5. Conclusion

In this exploration of music genre classification, the individual models, including K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN), were evaluated for their efficacy. Among these, the CNN exhibited superior accuracy, highlighting its proficiency in automatically extracting hierarchical features from spectrograms. While each model contributed uniquely to the classification task, CNN's dominance suggests the importance of deep

learning techniques in capturing complex patterns within the time-frequency domain of audio signals. Our future work is on exploring the potential benefits of combining the strengths of individual models through ensemble methods, such as stacking, bagging, or boosting, to create a more robust and accurate music genre classification system. To develop mechanisms for dynamic model adaptation that can adjust to evolving trends and diverse musical preferences, ensuring the system remains relevant over time.

References

1. Costa, Y.M., Oliveira, L.S., Koerich, A.L., Gouyon, F. and Martins, J.G., 2012, ―Music genre classification using LBP textural features‖ , ***Signal Processing***, Vol.***92***, No.11, pp.2723-2737.

2. Changsheng Xu, MC Maddage, and Xi Shao 2005, ―Automatic music classification and summarization. Speech and Audio Processing‖ , IEEE Transactions on, Vol.13, No.3, pp.441–450.

3. James Bergstra, Norman Casagrande, Dumitru Erhan, Douglas Eck, and Balázs Kégl 2006,―Aggregate features and adaboost for music classification‖ , Machine Learning, Vol.65, No.2-3, pp.473–484.

4. Roberto Basili, Alfredo Serafini, and Armando Stellato 2004 , ―Classification of musical genre: a machine learning approach‖, In ISMIR. Citeseer, 2004.

5. T Lidy, A Rauber, A Pertusa, and J Inesta 2007, ―Combining audio and symbolic descriptors for music classification from audio‖, Music Information Retrieval Information Exchange (MIREX).

6. Ioannis Panagakis, Emmanouil Benetos, and Constantine Kotropoulos, ―Music genre classification: A multilinear approach‖, In ISMIR, pages 583–588, 2008.

7. Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R Arce 2010, ―Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification‖, IEEE Transactions on Audio, Speech, and Language Processing,Vol. 18, No.3, pp.576–588.

8. George Tzanetakis and Perry Cook 2002, ―Musical genre classification of audio signals‖, IEEE transactions on Speech and Audio Processing, Vol.10, No.5, pp.:293–302.

9. Philippe Hamel, Simon Lemieux, Yoshua Bengio, and Douglas Eck 2011 , ―Temporal pooling and multiscale learning for automatic annotation and ranking of music audio‖, In ISMIR, pages 729–734.

10. Zehra Cataltepe, Yusuf Yaslan, and Abdullah Sonmez 2007, ―Music genre classification using midi and audio features‖, EURASIP Journal on Applied Signal Processing, Vol.2007, No.1, pp. 1-8.

11. Hrishikesh Deshpande, Rohit Singh, and Unjung Nam 2001, ―Classification of music signals in the visual domain‖, In Proceedings of the COST-G6 Conference on Digital Audio Effects, pp. 1–4.

12. Lambrou, T., Kudumakis, P., Speller, R., Sandler, M. and Linney, A., 1998, May. Classification of audio signals using statistical features on time and wavelet transform domains. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181) (Vol. 6, pp. 3621-3624). IEEE.