International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3084-3091 www.ijaem.net ISSN: 2395-5252

Music Genre Classification: A Comparison

G Mohan Krishna[1], Y Sujatha[2], R Vijay[3], S Akhila Yadav[4], P Meghnath Reddy[5]

[12345]*Department of Computer Science and Engineering*
Raghu Institute of Technology, Dakamarri, Visakahapatnam - 531162, AP, India.

Submitted: 10-07-2021 Revised: 23-07-2021 Accepted: 26-07-2021

ABSTRACT: Music has a significant impact on people's life. Music brings individuals together who share similar interests and serves as the glue that holds communities together. Communities can be identified by the type of music they write or listen to. Companies such as spotify,soundcloud uses music classification either to be able to place recommendations to the customers or simply as a product like shazam.Determining music genre is the first step in that process.Music genre classification automatically classifies different musical genres.This can be achieved by Machine Learning.Machine Learning techniques have proved to be quite successfully in extracting trends and patterns from a large pool of data.The same principles are applied in music analysis also.These Genres are created by humans.The goal of our study is to find a better machine learning algorithm that is superior to existing methods for predicting music genre.

Keywords: spectrograms; feature extraction; classification; K-nearest Neighbour (K-NN); Support Vector Machine (SVM); Convolution Neural Network(CNN); Mel Frequency Cepstral Coefficients (MFCC);

I. INTRODUCTION

Sound is represented as an audio signal with properties such as frequency, decibel, and bandwidth. The amplitude and time of a typical audio signal can be expressed as a function of each other. These audio signals are available in a variety of formats, making it possible for the computer to read and analyse them. The mp3 format, WMA (Windows Media Audio) format, and WAV (Waveform Audio File) format are only a few examples.

Music genre classification automatically classifies different musical genres. This project can be useful in identifying the genre of a song.This project is an application of Artificial Intelligence, more specifically Machine Learning, and also K-nearest algorithm that builds a system that predicts the genre of a song.In this project we are going to follow a simple approach to solve the classification problem and to draw comparisons with multiple other complex, robust models.We used 10 genres of music for the classification.

Music is differentiated by categorized classification known as genre. Humans are the ones who come up with these genres. A music genre is defined by the features that its members have in common. These features are usually linked to the music's rhythmic structure, instrumentation, and harmonic content.

Genre classification may be quite useful in explaining certain intriguing challenges, such as developing song references, hunting down related songs, and locating communities that will enjoy that specific music. It can also be applied for survey purposes in some cases.

II. LITERATURE SURVEY

Some recent publications reveal that research of musical genre in the context of music information retrieval are not yet complete, and they remain an active research issue, as seen by the following:

I. **In Hareesh Bahuleyan. Music Genre Classification using Machine Learning techniques**, their research presented a method for automatically classifying music by assigning tags to songs in the user's catalogue.They used both Neural Network and classic Machine Learning techniques in order to achieve their objectives.The first method employs a Convolutional Neural Network that is trained from start to finish utilizing the features of audio signal Spectrograms (pictures). The second method employs several Machine Learning techniques such as Logistic Regression, Random Forest, and others, as well as time and frequency domains from the audio signal.

II. **In Tom LH Li et al., Automatic musical pattern feature extraction using convolutional neural network.** They made an attempt to comprehend the key features that lead to the development of the best model for Music Genre Classification.Their strategy is to develop a robust model capable of identifying and segmenting an audio signal into speech, music, ambient sound, and quiet. This classification is broken down into two basic phases, making it useful for a variety of diverse uses.Classification between speech and non-speech is the initial step. A unique approach based on KNN (K-nearest-neighbor) and linear spectral pairs-vector quantization (LSP-VQ) has been devised in this paper.The second stage is to use a rule-based classification approach to separate the non-speech class into music, ambient noises, and quiet.

III. They have used a few uncommon and novel elements such as noise frame ratio and band periodicity, which are not only introduced but also studied in depth.A speech segmentation method was also incorporated and built.

IV. **In Lidy and Rauber**,the authors analyzed the role of psycho-acoustic aspects in distinguishing music genre, particularly the importance of STFT on the Bark Scale (Zwicker and Fastl, 1999) and (Tzanetakis and Cook, 2002) features such as mel-frequency cepstral coefficients (MFCCs), spectral contrast, and spectral roll-off are used.

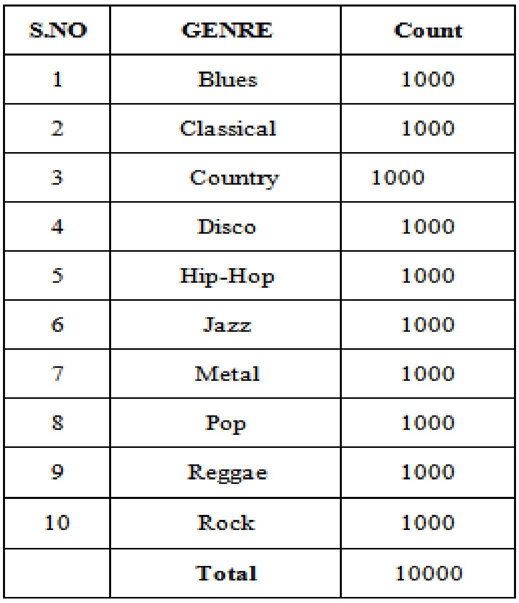
V. With deep neural networks' recent success, a number of researches have applied these approaches to speech and other types of audio data (AbdelHamid 2014; Gemmeke 2017). Due to the extreme high sampling rate of audio signals, representing audio in the time domain for neural network input is not easy.It has been handled by Van Den Oord (2016).

III. PROBLEM STATEMENT

The goal of this research is to build a machine learning model that categorize music into its genre and to compare the accuracies of this machine learning model to those of other models in order to make the appropriate conclusions.

IV. DATASET

In this project **GTZAN** – Music Genre Classification Data set is used.This GTZAN dataset consists of 10 music genres of 1000’s of audio files.It offers high-quality audio and pre-computed features, as well as metadata, tags, and free-form text such as biographies at the track and user levels.The number of music genres in the dataset has been tabulated in Table1.



S.NO	GENRE	Count
1	Blues	1000
2	Classical	1000
3	Country	1000
4	Disco	1000
5	Hip-Hop	1000
6	Jazz	1000
7	Metal	1000
8	Pop	1000
9	Reggae	1000
10	Rock	1000
	Total	10000

Table-1: Different Music genres

V. METHODOLOGY

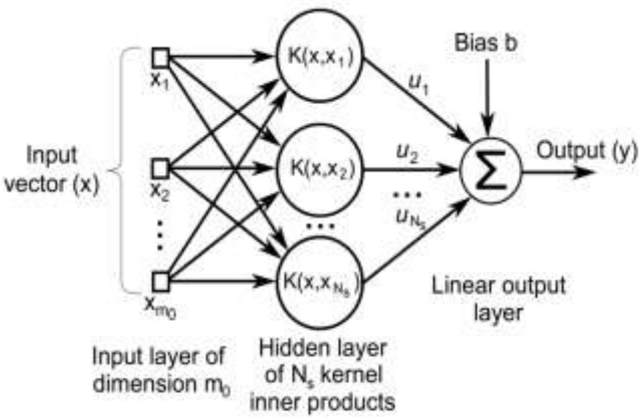
The specifics of data pre-processing are described in this part, followed by a discussion of the recommended strategy to this classification problem.

a) Support Vector Machine(SVM)

Support vector machines are a machine learning methodology that is based on the notion of structural risk minimization. In the field of pattern recognition, it has a wide range of applications. In order to estimate the decision function, SVM builds a linear model based on support vectors. SVM discovers the best hyper plane that separates the data without error if the training data are linearly separable**.**

SVM converts input patterns into higher-dimensional feature space via a non-linear mapping. A linear SVM is used to classify data sets that are linearly separable. The support vectors are the patterns that are maximum on the margins

The support vectors are (transformed) training patterns that are all near to the separation hyper plane. The most challenging patterns to categorize are the support vectors, which are training samples that form the ideal hyper plane. They're the most informative patterns in the categorization task, informally speaking.In the input space, the kernel function creates inner products to create machines with various forms of non-linear decision surfaces.

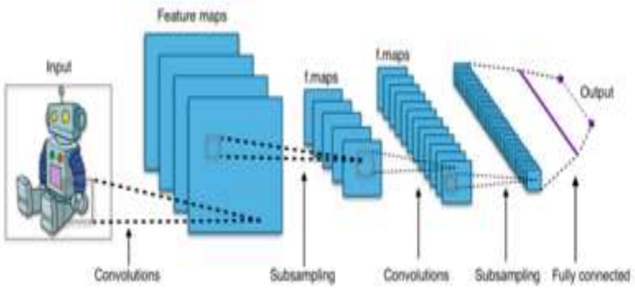


b) **Convolutional Neural Networks**

A CNN is a feed-forward network, which means that the input examples are sent back to the network and converted to an output.The outcome of supervised learning would be a label or a name applied to the data source.To put it another way, they map raw data to categories and detecting patterns that may indicate, for example, that an emergency situation exists

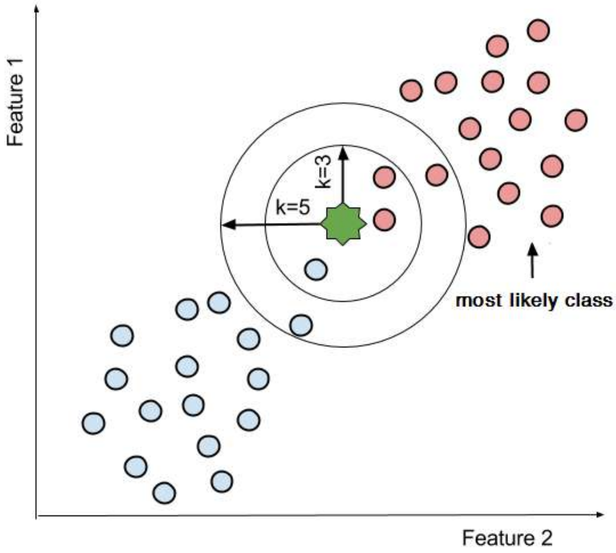
Label the input image as "folk" or "experimental."A feedforward network is trained on tagged pictures until it can estimate their categories with the least amount of error.

The network uses the trained set of parameters (collectively known as a model) to categorise input it has never seen before.A trained feedforward network can be subjected to any random collection of photos, and how it classifies the first shot will not necessarily change how it classifies the second. The net will not perceive a spectrogram of an experimental song after seeing a spectrogram of a traditional tune.That is, a feedforward network has no concept of temporal order and just considers the current example it has been exposed to as input. Feedforward networks have amnesia about their recent history, remembering only the early phases of training nostalgically.



c) **K- Nearest Neighbor-KNN:**

It is a distance-based supervised learning method. When using this approach to solve a classification issue, no model is generated, and the test operation is done on the labeled samples in the data set. The distance between the examples in the dataset will be used to create a new instance of the class label.The class tag is estimated from these determined distances by voting on the class labels of the nearest k.



VI. IMPLEMENTATION

I.Feature Extraction:

a)Time Domain Features:

The characteristics that were extracted from the raw audio stream are listed below.

- **Central moments:**The mean, standard deviation, skewness, and kurtosis of the signal's amplitude are the central moments.
- **Zero Crossing Rate (ZCR):** The signal changes sign from positive to negative at the zero crossing rate (ZCR). The amount of zero-crossings present in each frame is calculated when the complete 30-second signal is broken into smaller frames. As typical characteristics, the average and standard deviation of the ZCR over all frames were chosen.
- **Tempo:**The tempo of a piece of music refers to how quick or slow it is. The tempo is measured in BPM (beats per minute) (BPM). We use the Tempo's aggregate mean, which fluctuates from time to time.

II) Frequency Domain Features:

The Fourier Transform is used to convert the audio signal into the frequency domain. The following characteristics are then extracted.

- **Mel-Frequency Cepstral Coefficients (MFCC):** Davis and Mermelstein introduced MFCCs in the early 1990s, and they've shown to be quite beneficial for applications like speech recognition**.**
- **Chroma Features:** This is a vector that represents the signal's total energy in each of the 12 pitch classes. (C, C#, D, D#, E, F, F#, G, G#, A, A#, B).The mean and standard deviation are calculated using the sum of the Chroma vectors.

- **Spectral Contrast:** Each frame is separated into a number of frequency bands that are predetermined. The spectral contrast is determined as the difference between the greatest and minimum magnitudes within each frequency range.

The mean and standard deviation of the values collected across frames is regarded the representative final feature that is provided to the model for each of the spectral characteristics discussed above.

B. Classifier:

This section gives a quick rundown of the machine learning classifier used in this research.

- **Support Vector Machine:**SVM learning is a valuable statistic machine learning methodology that has been effectively utilized in the pattern identification domain**.**

Let's say we're given some training data(x1,x2,…,xn) and their class labels(y1,y2,…,yn) where x belongs to Rn and yi belongs to(-1,+1).and we'd like to divide the training data into two groups. The non-linear SVM classifier will be used if the data are linearly non-separable yet nonlinearly separable.

The core idea is to use non-linear transformation to turn input vectors into a high-dimensional feature space, and then conduct a linear separation in feature space.The inner product < x, y > is substituted by a kernel function K to create a non-linear SVM classifier (x, y).

- **Simple Artificial Neural Network (ANN)**:

A computational model based on the structure and functionality of biological neural networks is known as an artificial neuron network (ANN).

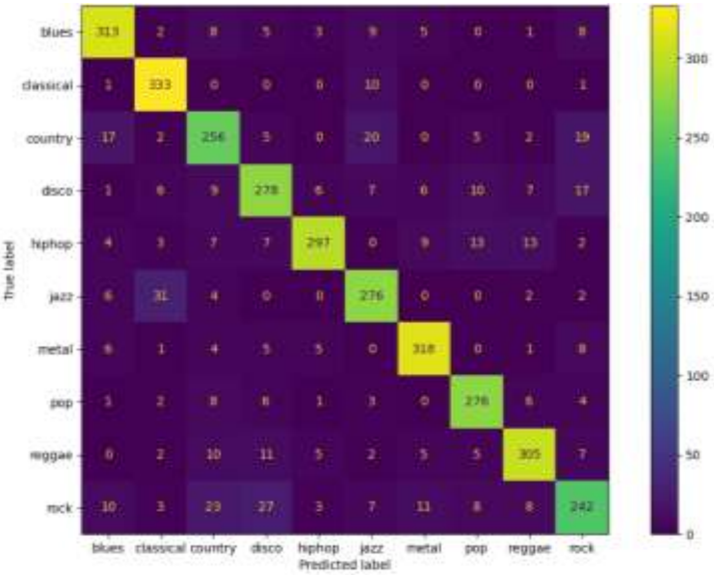
Because a neural network alters - or learns, in a sense - based on the input and output, the information that goes through the network modifies the structure of the ANN.ANNs are nonlinear statistical data modeling tools that are used to model or find patterns in complicated interactions between inputs and outputs. This model accepts a csv file containing handmade characteristics taken from audio clips.

- **K-Nearest Neighbour (KNN)**: K Nearest Neighbors is simple to implement a supervised learning algorithm that is widely used for classification. The simple idea of KNN is that similar items are near to each other, or in other words, the same traits exist nearby. The KNN classifier captures the notion of similarities among objects based on mathematics, like calculation of distance between the items.

In KNN, the test sample is assigned a class value to the class of the majority of its nearest neighbors. The KNN algorithm is based on the K value, which determines the number of training neighbors to which a test sample is compared. Here we took K value as 5.

VII. RESULTS

Here we are going to mention the Confusion Matrix and Accuracy details of methods SVM, CNN, K-NN. At first we used SVM (Support Vector Machine) for classification. Result is,
Confusion Matrix:

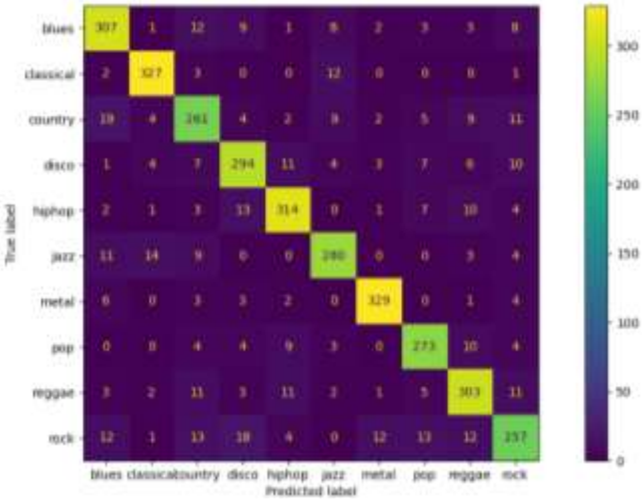


Classification Report:

	Precision	Recall	F1-score	support
Blues	0.88	0.88	0.88	354
Classical	0.86	0.97	0.91	345
Country	0.78	0.79	0.78	326
Disco	0.81	0.80	0.80	347
Hip-Hop	0.93	0.84	0.88	355
Jazz	0.83	0.86	0.84	321
Metal	0.90	0.91	0.91	348
Pop	0.87	0.90	0.88	307
Reggae	0.88	0.87	0.88	352
Rock	0.78	0.71	0.74	342

Accuracy			**0.85**	3397
Marco Average	0.85	0.85	0.85	3397
Weighted Average	0.85	0.85	0.85	3397

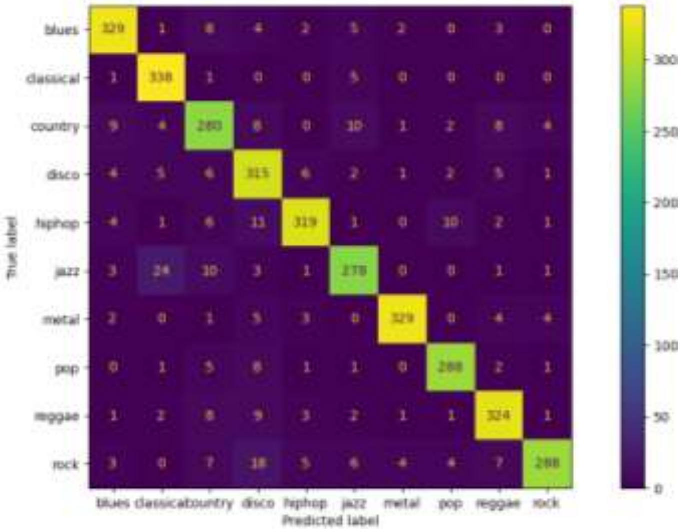
So, the SVM accuracy is **85%.**
Then, we used CNN (Convolution Neural Network) for classification.
Result is,
Confusion Matrix:



Classification:

	Precision	Recall	F1-score	support
Blues	0.85	0.87	0.86	354
Classical	0.92	0.95	0.94	345
Country	0.80	0.80	0.80	326
Disco	0.84	0.85	0.85	347
Hip-Hop	0.89	0.88	0.89	355
Jazz	0.88	0.87	0.88	321
Metal	0.94	0.95	0.94	348
Pop	0.87	0.89	0.88	307
Reggae	0.85	0.86	0.85	352
Rock	0.82	0.75	0.78	342
Accuracy			**0.87**	3397
Marco Average	0.87	0.87	0.87	3397
Weighted Average	0.87	0.87	0.87	3397

So, the CNN accuracy is **87%**.
Finally, we used K-NN (K-Nearest Neighbour) for classification.
Result is,
Confusion Matrix:

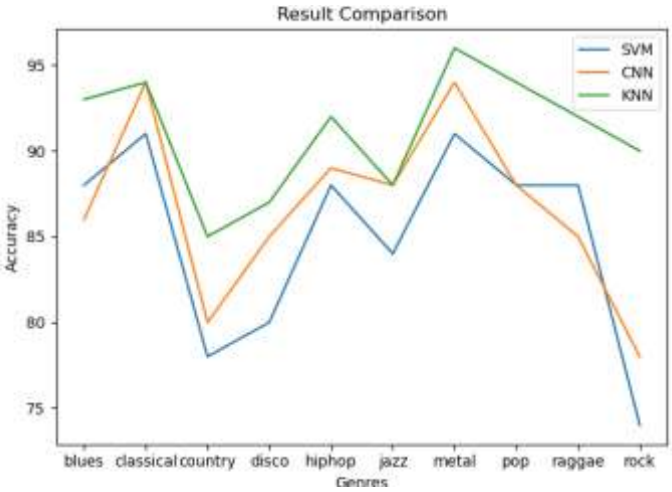


Classification Report:

	Precision	Recall	F1-score	support
Blues	0.92	0.93	0.93	354
Classical	0.90	0.98	0.94	345
Country	0.84	0.86	0.85	326
Disco	0.83	0.91	0.87	347
Hip-Hop	0.94	0.90	0.92	355
Jazz	0.90	0.87	0.88	321
Metal	0.97	0.95	0.96	348
Pop	0.94	0.94	0.94	307
Reggae	0.91	0.92	0.92	352
Rock	0.96	0.84	0.90	342
Accuracy			**0.91**	3397
Marco Average	0.91	0.91	0.91	3397
Weighted Average	0.91	0.91	0.91	3397

Better accuracy obtained by K-NN method is **91%**.

By the above results we come to know that first svm given lowest accuracy i.e., 85%,and then CNN came up with 87% and finally K-NN with 91%.



VIII. CONCLUSION

In this project GTZAN – Music Genre Classification Data set is used.We proposed a simple approach to solve the classification problem and we made comparisons with multiple other complex, robust models and increased some of the accuracy.K-Nearest Neighbour is determined to be the better model then k-Means, SVM, CNN under 10 types of music genres.

IX. FUTURE ENHANCEMENT:

We can further extend this Music Genre Classification project to map images of song i.e., cover images or genres in general.Probably, CNN can give good results for this work.

REFERENCES:

[1]. https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification
[2]. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
[3]. https://en.wikipedia.org/wiki/Melfrequency_cepstrum#:~:text=Mel%2Dfrequency%20cepstral%20coefficients%20(MFCCs,%2Da%2Dspectrum%22).&text=MFCCs%20are%20commonly%20derived%20as,windowed%20excerpt%20of)%20a%20signal.
[4]. https://pypi.org/
[5]. https://www.irjet.net/archives/V5/i10/IRJET-V5I10197.pdf
[6]. https://www.researchgate.net/publication/4015150_Musical_genre_classification_using_support_vector_machines
[7]. http://cs229.stanford.edu/proj2013/FauciCastSchulze-MusicGenreClassification.pdf
[8]. https://www.astesj.com/publications/ASTESJ_040221.pdf
[9]. https://towardsdatascience.com/musical-genre-classification-with-convolutional-neural-networks-ff04f9601a74
[10]. https://www.researchgate.net/publication/324218667_Music_Genre_Classification_using_Machine_Learning_Techniques
[11]. https://www.researchgate.net/figure/A-typical-example-of-a-KNN-classification-for-a-two-class-problem-ie-the-pink-and_fig2_322358139
[12]. https://en.wikipedia.org/wiki/Convolutional_neural_network
[13]. https://www.mdpi.com/1424-8220/14/11/20713
[14]. https://thesai.org/Publications/ViewPaper?Volume=8&Issue=8&Code=IJACSA&SerialNo=44
[15]. https://github.com/KishanMistri/Music-Genre-Classification
[16]. https://ieeexplore.ieee.org/document/1199998/