International Journal of Innovative Research in Engineering
Volume 3, Issue 3 (May-June 2022), PP: 513-517.
www.theijire.com
ISSN No: 2582-8746

Music Genre Classification Using Machine Learning

Pulkit Gigoo[1], Prof.Dnyaneshwar Kanade[2]
[1,2]Electronics and Communication Department, Vishwakarma Institute of Technology, Pune, Maharashtra, India.

How to cite this paper:
Pulkit Gigoo[1], Prof.Dnyaneshwar Kanade[2],
"Music Genre Classification Using Machine Learning",
IJIRE-V3I03-513-517.

Abstract: *Music plays an awfully necessary role in people's lives. Music brings similar folks along andis that the glue that holds communities along. Communities will be recognized by the sort of songsthat they compose, or even listen to. The aim of our project and analysis is to search out a higher machine learning formula than the pre-existing models that predict the genre of songs Therefore, it's essential toclassify the music in line with the genres to satisfy thewants of the folks flatly. The manual ranking of musicmay be a repetitive, protracted task and also the dutylies with the attender. The planned analysis work hascompared few classification models and established areplacement model for CNN, that is best thanantecedent planned models. This analysis work hastrained and compared the planned models onGTZAN dataset, wherever most of the models wereaudio file trains, whereas a couple of the models weretrained on theexposure.*

Key Word: *— deep leaning; music genre; GTZAN; convolution neural network; mel-spectogram*

I.INTRODUCTION

The increasing use of the Internet has wreaked havoc on the music industry, as well as triggering a slew of other developments. Every day, new songs are written in the music industry all over the world. Because categorizing such songs on a daily basis will become a tiresome task, technology can be utilized to heal the music and make classification easier or more efficient by utilizing its rhythms, beats, and lyrical composition. The audio signal can be used to represent a song. This audio signal has a variety of characteristics such as frequency, spectral roll-off, root- mean-square (RMS) level, bandwidth, zero-crossing rate, spectral center of mass, etc. firms (such as Soundcloud, Apple Music, Spotify, Wynk etc.) use music classification, either to put recommendations to their customers, or just as a product (like Shazam). To be ready to do any of the on top of 2 mentioned functions, deciding music genres is that the opening. To attain this, we are able to take the assistance of Machine Learning algorithms. These machine learning algorithms sway be terribly handy in Music Analysis too. Tremendous analysis has been tired genre classification. These researches are chiefly classified into 2 teams supported the sort of dataset used. 2 far-famed datasets area unit offered, they're FMA and GTZAN dataset. The machine learning approach is employed to resolve the challenges gift during this paper. Since, the introduction of the primary convolution neural network, it gave pace to the sphere of deep learning like image classification and segmentation, object detection and recognition. CNN could be a special reasonably neural network that incorporates a gird like topology. This grid will be a linear like statistic information or a 2nd gird like that of a picture. CNN uses a system like a multilayer perceptron that reduces the process needs. except for CNN, support vector machine, artificial neural network, multilayer perceptron and call tree area unit used for comparative analysis.

II. LITERATURE REVIEW

As the music business is growing speedily with new technological innovations, researchers also are developing their interest within the field of music. Varied machine learning solutions square measure projected by totally different authors to classify differing types of music.

I. In HareeshBahuleyan, (2018). expressive style Classification victimization Machine Learning techniques, the work conducted offers associate degree approach to classify music mechanically by providing tags to the songs gift within the user's library. It explores each Neural Network and ancient methodology of victimization Machine Learning algorithms and to realize their goal. the primary approach uses Convolutional Neural Network that is trained finish to finish victimization the options of Spectrograms (images) of the audio signal. The second approach uses numerous Machine Learning algorithms like supplying Regression, Random Forest etc, wherever it uses handmade options from time domain and frequency domain of the audio signal. The manually extracted options like Mel-Frequency Cepstral Coefficients (MFCC), color property options, Spectral center of mass etc area unit accustomed classify the music into its genres victimization cubic centimeter algorithms like supplying Regression, Random Forest, Gradient Boosting (XGB), Support Vector Machines (SVM). By examination the 2 approaches one by one they came to a conclusion that VGG-16 CNN model gave highest accuracy. By constructing ensemble classifier of VGG-16 CNN and XGB the optimized model with zero.894 accuracy was achieved.

II. In Tzanetakis G. et al., (2002). Musical style type of audio alerts, they've particularly explored approximately how the automated type of audio alerts right into a hierarchy of musical genres is to be finished. They consider that those track genres are express labels which are created through people simply to classify portions of track. They are labeled through a number of the not unusualplace traits. These traits are generally associated with the devices thatare used, the rhythmic structures, and primarily the harmonic track content material. Genre hierarchies are generally used to shape the very huge track collections that is to be had on web. They have proposed 3 characteristic sets: timbral texture, the rhythmic content material and the pitch content material. The research of proposed functions with a purpose to examine the overall performance and the relative significance changed into finished through schooling the statistical sample reputation classifiers through using a few actual-global audio collections. Here, on this paper, each entire report and the actual time frame-primarily based totally type schemes are described. Using the proposed characteristic sets, this version can classify nearly 61% of ten track style correctly.

III. In Lu L. et al., (2002). Content evaluation for audio category and segmentation, they've provided their have a look at of segmentation and category of audio content material evaluation. Here an audio movement is segmented in line with audio kind or speaker identity. Their method is to construct a strong version that is able to classifying and segmenting the given audio sign into speech, music, surroundings sound and silence. This category is processed in primary steps, which has made it appropriate for diverse different packages as well. In here, a singular set of rules that is primarily based totally on KNN (K- nearest-neighbors) and linear spectral pairs-vector quantization (LSP-VQ) is been evolved. The 2d step is to divide the non-speech elegance into music, environmental sounds, and silence with a rule- primarily based totally category method. Here they've made use of few uncommon and new capabilities together with noise body ratio, band periodicity which aren't simply introduced, however mentioned in detail. They have additionally protected and evolved a speaker segmentation set of rules. This is unsupervised. It makes use of a singular scheme primarily based totally on quasi - GMM and LSP correlation evaluation. Without any previous know-how of anything, the version can aid the open-set speaker, on-line speaker modelling and additionally the actual time segmentation.

IV. In Tom LH Li et al., (2010). Automatic musical sample function extraction the usage of convolutional neural network, they made an attempt to apprehend the principal functions which truly make contributions to construct the ultimate version for Music Genre Classification. The essential cause of this paper is to advise a unique technique to extract musical sample functions of the audio document the usage of Convolution Neural Network (CNN). Their center goal is to discover the opportunities of software of CNN in Music Information Retrieval (MIR). Their effects and experiments display that CNN has the sturdy ability to seize informative functions from the various musical sample. The functions extracted from the audio clips together with statistical spectral functions, rhythm and pitch are much less dependable and produces much less correct fashions. Hence, the technique made with the aid of using them to CNN, in which the musical records have comparable traits to photo records and specifically it calls for very much less previous knowledge. The dataset taken into consideration changed into GTZAN. It includes 10 genres with a hundred audio clips each. Each audio clip is 30 seconds, sampling charge 22050 Hz at sixteen bits. The musical styles have been evaluated the usage of WEKA device in which more than one class fashions have been taken into consideration. The classifier accuracy changed into 84 % and sooner or later were given higher. In evaluation to the MFCC, chroma, temp functions, the functions extracted with the aid of using CNN gave appropriate effects and changed into greater dependable. The accuracy can nevertheless be extended with the aid of using parallel computing on extraordinary aggregate of genres.

III. MATERIAL AND METHODS

A. Data set Preparation

Table No1. Number of Audio Clips in Each Genre

S. No	Class	Clips
1	Blues	100
2	Classical	100
3	Country	100
4	Disco	100
5	Hip-hop	100
6	Jazz	100
7	Metal	100
8	Pop	100
9	Reggae	100
10	Rock	100
	Total	1000

There are famous datasets to be had for Music Genre Classification, the FMA dataset which incorporates info of audio functions of 8000 exceptional songs of eight exceptional genres, the alternative is the GTZAN dataset which incorporates a thousand audio documents of 10 exceptional genres. The department of every style may be visible in Table I.

1) In this paper, the GTZAN dataset is used. This dataset has 10 classes, Blues, Classical, Country, Disco, Hip-Hop, Jazz, Pop, Metal, Reggae and Rock. Each magnificence containing a hundred audio clips, every clip being 30 seconds lengthy in .wav layout and is samples at 22050Hz, 16bit.

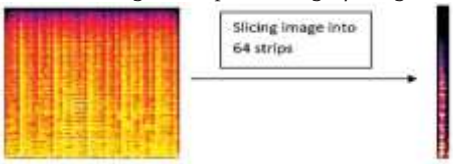
B. Data Pre-Processing

This dataset consists of one hundred audio clips for one elegance which isn't always sufficient to get an awesome accuracy. Either one ought to use a dataset with greater audio clips or pre-procedure the dataset in the sort of manner that it will increase the quantity of education and trying out samples.

1) Spectrogram Generation:

Each audio report is transformed into mel-spectrogram then audio clips are loaded the usage of the librosa library and generated the mel-spectrogram for every audio clip. After the spectrogram become generated, it become sliced into sixty-four strips. This accelerated our samples sixty-four times. We determined a complete of 64000 samples every of measurement 480x10 for schooling and testing. So, we divided out 64000 photos into for 44200 testing, 7000 for validation and 12800 for testing. Fig. 1 suggests the unique photo and the strips generated after processing.

Fig1. Pre-processing of images.



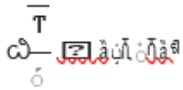
2) Feature Extraction:

Every audio may be represented in shape of an audio sign and this sign has exclusive capabilities. The audio capabilities are extracted which might be applicable for fixing the problem. These capabilities are divided into 2 sub categories. These definitions are stimulated through the paintings of writer in.

a) Time DomainFeature:

i. **Zero Crossing Rate**: The rate at which a signal changes from positive to negative or vice versa is known as zero-crossing rate.

ii. **Root Mean Square Energy***:* Root Mean Square Energy (RMSE) defines how loud a signal is. RMSE is defined as:



b) FrequencyDomain Features:

i. Mel-Frequency Capstral Coefficient: Set of features (round 10-20) that describe the form of the audio sign are referred to as Mel-Frequency Capstral Coefficient.

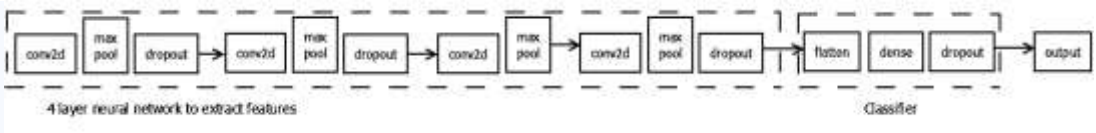
ii. Chroma Features: It is a illustration for a tune sign in which the complete spectrum is projected onto 12 bits representing the 12 wonderful semitones of musical octave.

iii. Spectral Centroid: It is the weighted imply of the frequencies gift withinside the sound, which tells us approximately the 'middle of mass'

3). **Convolution Neural Network:** As we are able to see there are traits capabilities even in a 480x10 length image, which can be distinct for each class. Our CNN version is supplied with those pix as input. A difficult structure of our version is proven in Fig. 2.

4) **CNN Model*:*** The schooling pix are exceeded i.e., the sliced pix of the spectrogram to our deep neural community for comprising of sub-networks. The first neural community is a four-layer convolution neural community for extracting capabilities from the pix. These extracted capabilities are then exceeded to the second one sub-community for classification. This community is absolutely related community containing absolutely related layers. In the stop a dense layer is used to expect the style of the audio. For satisfactory tuning and performance .

Fig2. CNN model



Each layer of the CNN does the following operations

a). Convolution: This layer has a fixed of filters whose parameters want to be learned. The size of every clear out out is smaller than the enter photo (normally 3x3 size). This clear out out passes thru the photo protecting all of the pixel values withinside the photo. As this clear out out passes thru the photo the scaler fabricated from the photo and the clear out out is calculated.

b) Max Pooling: The motive of pooling layer is to carry out down sampling consistent with the size of the given enter. Thus, lowering the wide variety of parameters with that activation. Two not unusualplace pooling techniques are common pooling and max pooling.

c) Dropout: It is a way to save you our version from overfitting thereby growing the efficiency. While schooling our version, in every new release the burden of a number of the neurons are set to 0 randomly. Final output is expected the usage of extraordinary mixture of neurons. A dropout fee of 0. four is used, i.e., a neuron weight is about to 0 for the duration of a new release, with a chance of 0. four.

Other Models*:* Here we talk in quick approximately the relaxation of our fashions that we designed. For this kind of fashions, we used a csv document which includes the handmade capabilities of the audio signals.

5) Other Models*:* Here we talk in quick approximately the relaxation of our fashions that we designed. For this kind of fashions, we used a csv document which includes the handmade capabilities of the audio signals.

A) Artificial Neural Network: It is a computational version this is stimulated through the running of a human brain. It is so due to the fact the data is travelled just like the human brain. As the data travels via the neurons of the community, the shape is affected because of which the neural community learns primarily based totally at the center and output. ANN is non- linear facts version, is used whilst complicated dating among the enter and out wishes to be found.

B) MLP*:* It is a sort of logistic regressor wherein we insert an intermediate layer additionally referred to as the hidden layer. Nonlinear activation function (tanh or sigmoid) is found in this layer. One could make the structure deep through putting as many hidden layers as consistent with the requirement of the user.

C) SVM*:* Tough SVM is commonly used for binary classification. Here we used one-vs-relaxation technique to finish our task. All the ten instructions are skilled one at a time and during FN for X is all of the X times that aren't categorized as X We have evaluated the overall performance of our fashions on the above-noted parameters, through thinking about the number of samples handed to the version. Table II offers tabulated data approximately the sensitivity, specificity and accuracy. The ROC curve in Fig. three suggests the real nice vs the false testing the class with the highest probability is selected as the predicted class.

D) Decision Tree: This classifier is used for multiclass classification. This kind of version may be pictured in shape of a binary tree. Starting from the foundation node to all of the inner nodes a hard and fast of inquiries to the dataset (associated with its features/attributes) is offered and the nodes are then similarly break up into new nodes having special characteristics. The leaves of the tree constitute the lessons wherein the dataset is break up.

IV. RESULTS

1) System Configuration and Dataset

Our proposed fashions are carried out on a PC having TensorFlow 1.12.0, tflearn 0. three.2, Keras 2.2.4, OpenCV three.4. three libraries hooked up in Python three.5.2 the use of an Intel Xeon three.4GHz processor and 32 GB RAM. For measuring the overall performance of our deep neural network, the GTZAN dataset is used. The dataset includes one thousand audio clips, one hundred audio clips for every of the magnificence. The time taken for our version turned into round 30 minutes.

2) Performance Measurement

The 3 parameters are used to estimate the overall performance of the version, sensitivity, specificity, and accuracy. Here sensitivity tells us how properly our fashions classify a specific magnificence, and specificity tells us how properly our version is classed for non-contemporary magnificence. Accuracy tells us approximately the general ratio of successfully detected events. All those parameters are described as follows:

Sensitivity = TP / (TP + FP) Specificity = TN / (TN + FP) Accuracy = (TP + TN) / Total Events

Here sensitivity and specificity are calculated for every magnificence and accuracy for the general results. For calculating sensitivity and specificity for a specific magnificence say X,

TP for X is all of the X times which might be categorized as X

TN for X is all of the non-X times that aren't categorized as X

FP for X is all the non-X instances that are classified as X

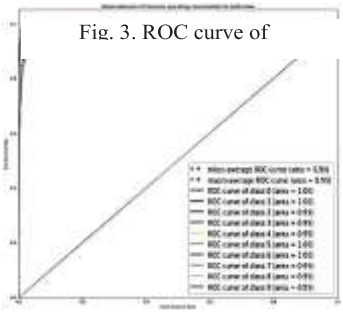


Fig. 3. ROC curve of

Since the AUC value of each class is great, our model is able to distinguish between each class.

Table No 2. Result Analysis of Our Model

Class	Sensitivity	Specificity	Accuracy
Blues	**0.93**	**0.93**	**0.93**
Classical	**0.92**	**0.98**	**0.95**
Country	**0.87**	**0.87**	**0.87**
Disco	**0.90**	**0.85**	**0.88**
Hiphop	**0.91**	**0.89**	**0.90**
Jazz	**0.91**	**0.93**	**0.92**
Me tal	**0.93**	**0.97**	**0.95**
Pop	**0.91**	**0.90**	**0.90**
Reggae	**0.91**	**0.86**	**0.89**
Rock	**0.86**	**0.86**	**0.86**
OverallAccuracy			**0.91**

3). Comparative Analysis Measurement

In this paper, we've got designed 5 one-of-a-kind fashions CNN, ANN, SVM, MLP and Decision Tree. The following one-of-a-kind audio features (Mel-Frequency Capstral Coefficient, Root Mean Square Energy, Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, Spectral Roll-off) are saved in a csv record and surpassed it to our ANN, SVM, MLP and Decision Tree fashions. For our proposed CNN version, we've got surpassed the extracted mel-spectrogram images. The accuracy completed with the aid of using our version is given in Table III. Fig. four indicates the end result evaluation of our one-of-a-kind fashions.

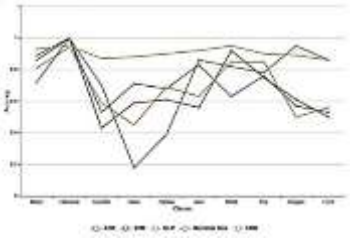


Table No. 3 Accuracy of Different Models

Mode l	Accuracy
ANN	70%
SVM	68.9%
MLP	68.7%
Decision Tree	74.3%
CNN	91%

It may be visible from the graph that our CNN version and ANN version gave the excellent results. Out of the ten genres our fashions gave excellent accuracy for blues, classical, pop. Overall, the accuracy accomplished via way of means of every

version is given withinside the Table III. Our CNN version had gain on disco and hip-hop genre. It may be visible all our version relatively finished worst on united states genre. The confusion matrix of our CNN version is proven in Table IV.

There are lot of techniques proposed on these paintings of literature. People have used unique varieties of dataset and a few authors created their personal dataset. Moreover, a few the numbers of genres taken into consideration via way of means of the authors are unique in unique works. Some authors have taken into consideration five or much less out of the ten genres for constructing their models. Hence, their effects cannot be in comparison with this version. We have handiest taken into consideration handiest the ones work wherein all of the 10 genres of the GTZAN dataset are used. Table V demonstrates our version via way of means of evaluating it with via way of means of evaluating the overall performance with different pre-present models.

Table No. 4 *Comparative Analysis with other Models*

Model Used	Accuracy %
Residual Neural Network	94.0
Compressive Sampling	92.7
Convolution Neural Network	90.7
Convolution Neural Network	88.5
Support Vector Machine	84.4
AM-FM	84.3
Convolution Neural Network	65
Residual Neural Network	64
Convolution Neural Network	91

V.CONCLUSION

The proposed studies paintings has applied the GTZAN dataset and produced more than one fashions to finish this undertaking on this piece of track classification. The proposed version has used more than one inputs for numerous fashions in conjunction with the audio mel- spectrogram and transferred this to our CNN, and numerous sound record traits saved withinside the ANN, SVM, MLP and Decision Tree csv information 91%, equal to the human know-how of style with maximum correct achievement. Since, a few patterns have been pretty different and a few alternatively different together with the u. s. and the rock style have been stressed with different patterns, even though conventional and blues have been without problems identified.

References

[1] *McKinney,Martin,andJeroenBreebaart. "Features for audio and music classification."(2003).*

[2] O'Shea,Keiron,andRyanNash."An introduction to convolutionalneural networks." arXiv preprint arXiv:1511.08458(2015).

[3] Bisharad,Dipjyoti,andRabulHussainLaskar."Music Genre Recognition Using Residual Neural Networks."InTENCON2019-2019IEEERegion10 Conference (TENCON), pp. 2063-2068. IEEE,2019.

[4] Zhang, Scott, HuapingGu, and Rongbin Li. "MUSIC GENRE CLASSIFICATION: NEAR-REALTIME VSSEQUENTIAL APPROACH." (2019).

[5] Chillara, Snigdha, A. S. Kavitha, Shwetha A. Neginhal, Shreya Haldia, andK.S.Vidyullatha."MusicGenreClassification usingMachineLearning Algorithms: A comparison."(2019).

[6] Bahuleyan,Hareesh."Musicgenreclassificationusingmachinelearning techniques." arXiv preprint arXiv:1804.01149(2018).

[7] Yang,Hansi,andWei-QiangZhang."MusicGenreClassificationUsing DuplicatedConvolutionalLayersinNeuralNetworks."InINTERSPEECH, pp. 3382-3386. 2019.

[8] Gessle, Gabriel, and Simon Åkesson. "A comparative analysis of CNN and LSTM for music genre classification."(2019).

[9] Defferrard, Michaël, KirellBenzi, Pierre Vandergheynst, and Xavier Bresson. "Fma: A dataset for music analysis." arXiv preprint arXiv:1612.01840(2016).

[10] George,Tzanetakis,EsslGeorg,andCookPerry."Automaticmusical genreclassificationofaudiosignals."InProceedingsofthe2ndinternational symposium on music information retrieval, Indiana .2001.

[11] Srivastava,Nitish,GeoffreyHinton,AlexKrizhevsky, IlyaSutskever, andRuslanSalakhutdinov."Dropout:asimplewaytopreventneuralnetworks fromoverfitting."Thejournalofmachinelearningresearch15,no.1(2014): 1929-1958.

[12] Grossi,Enzo,andMassimoBuscema."Introduction toartificial neural networks."Europeanjournalofgastroenterology&hepatology19,no.12 (2007):1046-1054.

[13] Weston,Jason,andChrisWatkins."Supportvectormachinesformulti- class pattern recognition." In Esann, vol. 99, pp. 219-224.1999.

[14] Chang, Kaichun K., Jyh-Shing Roger Jang, and Costas S. Iliopoulos. "Music Genre Classification via Compressive Sampling." In ISMIR,pp.387- 392. 2010.

[15] Hamel,Philippe,andDouglasEck."Learningfeaturesfrommusicaudio with deep belief networks." In ISMIR, vol. 10, pp. 339-344.2010.

[16] *Zlatintsi, Athanasia, and Petros Maragos. "Comparison of different representations based on nonlinear features for music genre classification." In201422ndEuropeanSignalProcessingConference(EUSIPCO),pp.1547- 1551. IEEE,2014.*