JOURNAL OF CRITICAL REVIEWS

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

MACHINE LEARNING CONCEPT BASED ALGORITHMS FOR MUSIC GENRE CLASSIFICATION SYSTEM

C. Karthik[1]*, K. Vinothkumar[2], S. Vigneshwaran[3]

[1]*Assistant professor, saraswathy college of Engineering and technology, Tindivanam
[2]Assistant professor, VRS college of Engineering and technology, viluppuram
[3]Assistant professor, sri vidya college of Engineering & Technology, viruthunagar

***Corresponding Author:** C. Karthik
*Assistant professor, saraswathy college of Engineering and technology, Tindivanam

Abstract:
Music is like a mirror, and it tells people a lot about who you are and what you care about, whether you like it or not. Music can be classified into taxonomies based on genre, performer, composer or geographic or cultural point of origin. Music genres can be seen as categorical descriptions used to segregate music based on various characteristics such as instrumentation, pitch, rhythmic structure, and harmonic contents. The top 10 genres in the music industry are blues, classical, country, disco, hip-hop, jazz, reggae, rock, metal and pop. Automatic music classification is an area of research that has been receiving a great deal of attention in recent years due to the rapid growth of digital entertainment industry. There are two major challenges with music genre classification: Firstly, musical genres are loosely defined, so that people often argue over the genre of a song. Secondly, extracting differentiating features from audio data that could be fed to the model is a nontrivial task. Although music genre classification has been a challenging task in the field of Music Information Retrieval (MIR), automatic music genre classification is important for music retrieval in large music collections on the web. This project aims to build a machine learning classifier after scrutinizing various machine learning algorithms that classifies music based on its genres. The chosen classifier, Support Vector Machines then learns from the data, explores the performance of various features extracted from the audio signal and classifies the genre of the audio input. This project can be extended to develop various systems like music genre-based disco lights and emotion-mapped music systems.

Keywords: Rhythmic structure, harmonic contents, genres, Automatic music genre classification, Music Information Retrieval, music collections, machine learning classifier, machine learning algorithms, Support Vector Machines, music genre-based disco lights, emotion-mapped music systems.

1. Introduction

The rapid development of various affordable technologies for multimedia content capturing, data storage, high bandwidth in speed transmission and the multimedia compression standards such as JPEG and MPEG, have resulted in a rapid increase of the size of digital multimedia data collections and greatly increased the availability of multimedia contents to the general user. Digital music is one of the most important data types distributed by the Internet and the amount of digital music increases rapidly nowadays. Music classification is an interesting problem with many applications, from Drinkify (a program that generates cocktails to match the music) to Pandora to dynamically generating images that complement the music

However, music genre classification has been a challenging task in the field of music information retrieval (MIR). Music genres are hard to systematically and consistently describe due to their inherent subjective nature. The main motivation of this system is to reduce the time-spent in order to classify music and reduce the human-made errors while classifying music. The goal of the project is to train a system that can automatically classify the audio file submitted by the user into any one of the categories of genre. Music plays a very important role in people's lives. Music brings like-minded people together and is the glue that holds communities together. Communities can be recognized by the type of songs that they compose, or even listen to. Different communities and groups listen to different kinds of music. One main feature that separates one kind of music from another is the genre of the music. Hence the best algorithm which has the highest accuracy will be chosen and then it will be coded in such a way to classify the audio sample efficiently.

The objectives of the proposed system are as follows:

1. To build a machine learning model that classifies music into genres based on various different features, instead of manually entering the genre.
2. To reach a good accuracy so that the model classifies new music into its genre correctly
3. This model should be better than at least a few pre-existing models

Supervised learning is a powerful tool for data classification. Supervised learning is a learning model that was built to predict the given unforeseen input instance. With supervised learning, the labelled data can be used, which is the data that has been classified to infer a learning algorithm. A supervised learning algorithm takes a known set of input data

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

(the learning set) and known responses to the data (the output) to learn a classification model. A learning algorithm then trains a model to predict a new instance of data. Supervised algorithms use classification and regression techniques to develop predictive models. There are a number of approaches that have been proposed in the past for music genre classification system. The objective of this survey is to do a comparative study in order to plump for a classification algorithm that suits best for music genre classification. This literature survey contains papers on Music Genre Classification System using several algorithms.

2. Literature Survey

Meimei et.al. [1] attempts to classify the genre of music using a Double Weighted KNN algorithm (DW-KNN). This algorithm makes improvements in two aspects, both in distance calculation and category judgement of traditional KNN algorithm. The first step in music genre classification is dataset collection. The second step is data preprocessing which includes generating characteristic matrix by extracting 59-dimensional characteristic of each songs; so, the result is to obtain a characteristic matrix with 59 columns and 1000 rows, forming the training set and the test set. The dataset is divided into four equal parts where three parts are used as the training set and one part as the test set. Then it performs loop test and data normalization. The data in sample set is normalized uniformly into the range of [-1,1]. The third step is evaluation criteria and validation methods. The cross-validation method divides the sample randomly into k collections. It selects k-1 collections as training set and remaining as testing set. The final step is observing the experimental results and analysis, which involves selecting the value of k. This algorithm has better classification accuracy rate for mass music data classification.

Liang et.al. [2] proposed a transfer learning approach for audio-based classification of 11 western music genres. The dataset used here consists of 100 audio tracks for each genre, having 1100 in all in mp3 format, out of which 75% is used as training data and remaining is used as testing data. Transfer learning can be applied for different classification and regression tasks. The results show that the system does not separate pop from other genres. So, the future work involves fine tuning the system to classify pop genre.

Ghosal et.al. [3] propounded an automatic music genre classification system using a deep learning model. The dataset used here is GTZAN dataset, that contains 10 music genres (blues, classical, country, disco, hip-hop, pop, jazz, reggae, rock, metal), each with 100 audio clips in .au format. The dataset incorporates samples from a variety of sources like CDs, radios, microphone recordings, etc. The training, testing and validating sets are randomly partitioned following the proportion 8:1:1. The model consists of a four-layer convolutional neural network (CNN) of 64 feature map, 3-by-3 convolutional kernels and max pooling layers. The output of CNN is a sequence in which every timestamp relies on the immediate predecessors and long-term structure of the entire song. LSTM Sequence to Sequence Autoencoder is used to capture both transient and overall characteristics and to learn representations of time series data by taking into account their temporal dynamics. Clustering Augmented Learning Method (CALM) classifier is used for classification. CALM is based on the concept of simultaneous heterogeneous clustering and classification to learn deep feature representations of the features obtained from LSTM autoencoder. Computational Experiments using GTZAN dataset resulted in an overall test accuracy of 90% with a precision of 85%. The future involves improvising new distance metric methods to compute the similarity between genres.

Vishnupriya et.al. [4] proposed automatic music genre classification using Convolution Neural Network (CNN). Music genre classification involves feature extraction and classification. Initially features are extracted from the waveform later using these features classifier is built for training. The dataset used contains 10 genres namely blues, classic, country, disco, rap, jazz, heavy metal, popular, reggae and rock. Each genre contains 100 recording, making it 1000 in all. The features are then extracted from the music. The accuracy of the model is calculated using

$$\text{Accuracy} = \frac{\text{No of songs correctly classified X 100}}{\text{Total number of songs}}$$

Jawaherlalnehru et.al. [5] presents a comprehensive machine learning approach to the problem of automatic musical genre classification of audio signal. The first step is data collection. The music database consists of 400 audio tracks with metadata. For each genre (considering only 4 genres namely classical, pop, rock, and electronic) 100 audio tracks of 60sec long are considered. All the audio files are in .au format with 44.1 KHz sampling frequency, stereo and 16bit PCM. the dataset is partitioned randomly into three parts: 60% for training, 20% for validation, 20% for testing. The next step, feature extraction is the process that converts an audio signal into a sequence of feature vectors. Feature extraction reduces the redundant information from audio signal and provides a compact representation. The most popular technique Mel Frequency Cepstral Coefficients (MFCC) is used. It is based on a linear cosine transform a log power spectrum on a nonlinear mel scale of frequency. Each short-term Fourier transform magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated. Then discrete cosine transform is applied to the log of the mel spectral coefficients to obtain MFCC. The final step, classification is the process by which particular label is assigned to a particular music format. The system is developed using a Multi label feed-forward Deep Neural Network (DNN) to recognize the genres. The DNN is fully connected neural network which consists of one input layer, one output layer, and several hidden layers. The number of neurons of input depends on the dimensions of input feature vectors, while the number of neurons of output layers is equal to the number of music genres being considered. The

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

proposed system observed higher classification accuracy of 97.8%. The future work includes increasing databases and using other feature techniques and to find an effective method to combine ensemble method with DNN architecture

Quinto et.al. [6] proposed designing of a classifier for Jazz sub-genre classification using an LSTM. The dataset of three sub-genres of jazz which are swing, bebop and acid jazz are under concern. The accuracy of this system was about 90%. The future works would be training the classifier to classify up to 10 genres from GTZAN dataset, adding custom penalty matrix for misclassification and implementing larger network configurations such as increasing the number of LSTM units and the number of layers and finally trying to replace MLP with CNN.

Weibin et.al. [7] proposed two ways to improve music genre classification with convolutional neural networks. The two ways are: 1) combining max-pooling and average-pooling to provide more statistical information to higher level neural networks; 2) using shortcut connections to skip one or more layers, a method inspired by residual learning method. The dataset used is GTZAN dataset, which was collected by Tzanetakis and Cook. There are 1000 song excerpts that are almost evenly distributed into ten different genres: Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae and Rock. Each song excerpt lasts about 30 seconds and is sampled at 22050Hz, 16 bits; as this improved the classification accuracy. The input of the CNN is simply the short time Fourier transforms of the audio signal. The output of the CNN is fed into another deep neural network to do classification. The output of the networks are the probabilities of different genres for each music clip. The probabilities of the clips from the same song are added and the genre with the maximum value is chosen as the label of the song. By comparing two different network topologies, preliminary experimental results on the data set show that the above two methods can effectively improve the classification accuracy, especially the second one. The future enhancement of this project is to fuse new methods such as multi-scale convolution and pooling with residual learning and study end-to-end learning to extract salient musical representations from the raw audio signals directly.

Jeong et.al. [8] describes a framework for temporal feature learning from audio with normalized cepstral modulation spectrum and deep neural network and applies it to classify music genres. The dataset used is the GTZAN dataset, that consists of 1000 (100 from each of 10 genres) 30- second long music clips with the sampling rate of 22050Hz. The results were examined by partitioning the dataset in two ways. 1. Dividing the dataset randomly into three groups (50% for training, 25% for validation and 25% for testing) and performing the experiments four times to present the averaged results. Since this random partitioning cannot be trusted, they moved for the next method. 2. "Fault-filtered" partitioning, where the dataset is divided into 443/197/290 to avoid repetition of artist across training, validation and testing sets. DNN was trained using mini- batch gradient descent with 0.01 step size and 100 batch size for the proposed algorithm. Optimization procedure was done after 200 epochs. Genre classification was performed using random forest (RF) with 500 trees as a classifier. Each music clip of 30s was first divided into a number of 5s-long short segments with 2.5s overlap. Classification was performed on each 5s- long segment, and used majority voting to classify the whole music clip. It is noted that both training and validation data were used to train RF since it does not require additional data for validation. To inspect the performance of different features, using random and fault-filtering partitioning, the features from test data were visualized using a 2-dimensional projection. The overall accuracy of this proposed system is 85%. The classification accuracies are higher with random partitioning because of artist repetition. The future enhancement is to apply the proposed method to various MIR related tasks, mood classification and instrument identification.

Rajanna et.al. [9] involves a two-layer neural network with manifold learning techniques for music genre classification. Preprocessing and extraction of meaningful audio features was difficult and challenging task of music classification. Adding to it, appropriate choice of a learning model is next challenging task. Input is a set of raw audio signals which needs to be processed. This proposed method involves preprocessing of audio signals, feature extraction, dimensionality reduction techniques, classification model such as SVMs or DNNs and genre label prediction of test samples. Future enhancements include the study other network architectures such as convolutional neural networks and stacked autoencoders for music classification and explore different signal preprocessing and representations to measure the network sensitivity for classification.

Haggblade et.at. [10] investigates various machine learning algorithms including k-nearest neighbor (k-NN), k-means, multi-class SVM and neural networks to classify music genre. The dataset used here is GTZAN Genre collection containing 1000 audio tracks each 30 seconds (100 tracks in each genre). In this paper, the 4 distinct genres: classical, jazz, metal and pop are classified. Thus, the total dataset contains 400 songs, of which 70% is used for training and 30% is used for testing and measuring results. For audio processing, Mel Frequency Cepstral coefficients (MFCC) is used. The fundamental calculation in k-NN training is to figure out the distance between two songs. This is computed using Kullback-Leibler divergence.

$$\mathbf{D_{KL}\,(p\,,\,q) = KL(p||q) + KL(q||p)}$$

The first ML technique used to classify genre is k-nearest neighbors (k-NN) which is known for its ease of implementation. For unsupervised k-means clustering to work on the feature set, a custom implementation was written to determine how to represent cluster centroids and how to update centroids in each iteration. SVM classifiers provide a reliable and fast way to differentiate between data with only two classes. In order to generalize SVMs to data falling into multiple classes (i.e. genres), directed acyclic graph (DAG) of two-class SVMs trained on each pair of class is used. The next ML technique is neural network. The data is randomly split by a ratio of 70:15:15- 70% of the data for training the

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

neural network, 15% of the data for verification to ensure that there is no over-fitting, and 15% of the data for testing. After multiple test runs, the feedforward model with 10 layers for neural network model gives the best classification results. The overall accuracy of k-NN and SVM is 80% and 87%. This work doesn't give a completely fair comparison between learning techniques for music genre classification. The future work includes adding a validation step to the DAG SVM would help determine which learning technique is superior in this application. In addition, including additional metadata text features such as album, song title, or lyrics could allow us to extend this to music mood classification.

Yaslan et.al. [11] had 1000 music and each music having maximum of 30 seconds in length and the genres present in the data set are classical, country, disco, hiphop, jazz, rock, blues, reggae, pop, metal. They have experimented with ten different classifiers like KNN, Naïve bayes classifier, etc. They have used forward and backward selection algorithm to select best feature activity and they have used Principal Component Analysis to reduce the dimensionality of feature set. They have combined some classifier for improving the accuracy of classification and concluded that the classification is more accurate when classifiers are combined for classification.

Alessandro et.al. [12] suggested to extract features from various parts rather from single part so that performance of classification is improved. They have designed multilayer perceptron neural network classifier with one hidden layer. They have combined the outputs of the three neural networks to compensate the drawbacks of each neural network. Improved musical genre classification is achieved when the two best single classifiers are combined through the weighted sum or weighted product rule.

Meng et.al. [13] focused on extracted features based on three time scales short-term. They have introduced a feature integration technique called AR model, has been proposed as an alternative to the dominating mean-variance feature integration. They have used two classifiers a single layer neural network and a Gaussian classifier based on the covariance matrix for the purpose of classification.

Li et.al. [14] presented a comparative study between the features included in the MARSYAS framework and a set of features based on Daubechies Wavelet Coefficient Histograms (DWCH), using also other classification methods such as SVM and Linear Discriminant Analysis. For comparing they have employed two datasets. One dataset with features extracted from the beginning of the music signal, and other one being dataset composed by 755 music pieces having five music genres, with features extracted from the interval that goes from second 31 to second 61. Experimental results prove that the SVM classifier has more accuracy than other methods in case of first dataset it improves accuracy to 72% using the original feature set and to 78% using the DWCH feature set, in the second dataset results were 71% for the MARSYAS feature set and 74% to the DWCH feature set.

Tzanetakis et.al. [15] proposed set of features to represent a music piece and those features were obtained from a signal processing perspective which includes pitch related features, beat-related features and timbral texture features and the features were extracted from the first 30- seconds of each music. For classification they have used Gaussian classifiers, Gaussian mixture models and the k Nearest-Neighbors classifier. Their dataset had one thousand samples containing ten music genres. Obtained results indicate an accuracy of about 60%.

Hareesh Bahuleyan [16] gives an approach to classify music automatically by providing tags to the songs present in the user's library. It explores both Neural Network and traditional method of using Machine Learning algorithms and to achieve their goal. The first approach uses Convolutional Neural Network which is trained end to end using the features of Spectrograms (images) of the audio signal. The second approach uses various Machine Learning algorithms like Logistic Regression, Random forest etc, where it uses hand-crafted features from time domain and frequency domain of the audio signal. The manually extracted features like Mel- Frequency Cepstral Coefficients (MFCC), Chroma Features, Spectral Centroid etc are used to classify the music into its genres using ML algorithms like Logistic Regression, Random Forest, Gradient Boosting (XGB), Support Vector Machines (SVM). By comparing the two approaches separately they came to a conclusion that VGG-16 CNN model gave highest accuracy. By constructing ensemble classifier of VGG-16 CNN and XGB the optimised model with 0.894 accuracy was achieved

Tom LH Li et.al. [17] made an effort to understand the main features which actually contribute to build the optimal model for Music Genre Classification. The main purpose of this paper is to propose a novel approach to extract musical pattern features of the audio file using Convolution Neural Network (CNN). Their core objective is to explore the possibilities of application of CNN in Music Information Retrieval (MIR). Their results and experiments show that CNN has the strong capacity to capture informative features from the varying musical pattern. The features extracted from the audio clips such as statistical spectral features, rhythm and pitch are less reliable and produces less accurate models. Hence, the approach made by them to CNN, where the musical data have similar characteristics to image data and mainly it requires very less prior knowledge. The dataset considered was GTZAN. It consists of 10 genres with 100 audio clips each. Each audio clip is 30 seconds, sampling rate 22050 Hz at 16 bits. The musical patterns were evaluated using WEKA tool where multiple classification models were considered. The classifier accuracy was 84 % and eventually got higher. In comparison to the MFCC, chroma, temp features, the features extracted by CNN gave good results and was more reliable. The accuracy can still be increased by parallel computing on different combination of genres

Paradzinets et.al. [18] explored acoustic information, beat-related and timbre characteristics. To obtain acoustic information they used Piecewise Gaussian Modeling (PGM) features enhanced by modeling of human auditory filter. To

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

do so, they obtained the PGM features, then applied critical bands filter, equal loudness and specific loudness sensation. To extract the beat-related characteristics, they used wavelet transforms, getting the 2D-beat histograms. For the timbre characteristics, they collected all detected notes with relative amplitude of their harmonics and then computed their histograms. Among others issues, their results show: (i) an improvement when using perceptually motivated PGM instead of basic PGM, i.e., accuracy of 43% versus 40.6%; (ii) training different NNs for each genre is better than training only one NN with all the genres being considered, which corresponds to an average accuracy of 49.3%.

Jang et.al. [19] proposed a novel music genre classification system based on two novel features and a weighted voting. The proposed features, modulation spectral flatness measure (MSFM) and modulation spectral crest measure (MSCM), represent the time-varying behavior of a music and indicate the beat strength. The weighted voting method determines the music genre by summarizing the classification results of consecutive time segments. Experimental results show that the proposed features give more accurate classification results when combined with traditional features than the octave-based modulation spectral contrast (OMSC) does in spite of short feature vector and that the weighted voting is more effective than statistical method and majority voting

Lu L. et.al. [20] presented their study of segmentation and classification of audio content analysis. Here an audio stream is segmented according to audio type or speaker identity. Their approach is to build a robust model which is capable of classifying and segmenting the given audio signal into speech, music, environment sound and silence. This classification is processed in two major steps, which has made it suitable for various other applications as well. The first step is speech and non- speech discrimination. In here, a novel algorithm which is based on KNN (K- nearest- neighbour) and linear spectral pairs-vector quantization (LSP-VQ) is been developed. The second step is to divide the non-speech class into music, environmental sounds, and silence with a rule- based classification method. Here they have made use of few rare and new features such as noise frame ratio, band periodicity which are not just introduced, but discussed in detail. They have also included and developed a speaker segmentation algorithm. This is unsupervised. It uses a novel scheme based on quasi - GMM and LSP correlation analysis. Without any prior knowledge of anything, the model can support the open-set speaker, online speaker modelling and also the real time segmentation.

This literature survey discusses in detail all advances in the area of music genre classification. The most accurate solution provided in this area directly or indirectly depends upon the quality, pros and accuracy provided by the method. Various techniques have been described in this paper for the classification of music genres. The comparison table shown below delineates the differences between the algorithms proposed in the existing systems. From the study done so far, it has been analysed that the selection of the algorithm plays a crucial role in order to attain good rate of accuracy. Studies in the paper reveals that there is still scope to enhance the algorithms as well as to intensify the accuracy rate.

This survey was mainly done for choosing an algorithm that suites best in implementing a music genre classification system. The best algorithm is chosen based on the accuracy and the efficient working of the system. The results of this literature survey show that K-nearest neighbors and Support Vector Machine suits well for the problem statement. Various researches have shown that KNN gives best results for this problem. K-Nearest Neighbors is a popular machine learning algorithm for regression and classification. It makes predictions on data points based on their similarity measures i.e., distance between them. Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression. SVM scales relatively well to high dimensional data. Taking these advantages on consideration, the KNN and SVM are chosen as the best algorithm for music genre classification system.

3. System Design

System Design is the process of defining the architecture for a system to satisfy the specified requirements. System design is the process of designing the elements of the system such as the architecture, modules and the components of the system, the different interfaces of those components and the data that goes through the system.

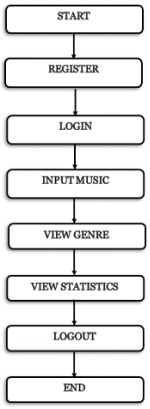


Figure 1 Block Diagram (User)

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

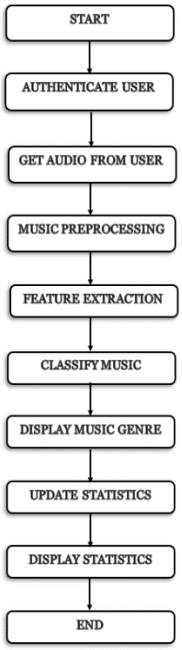
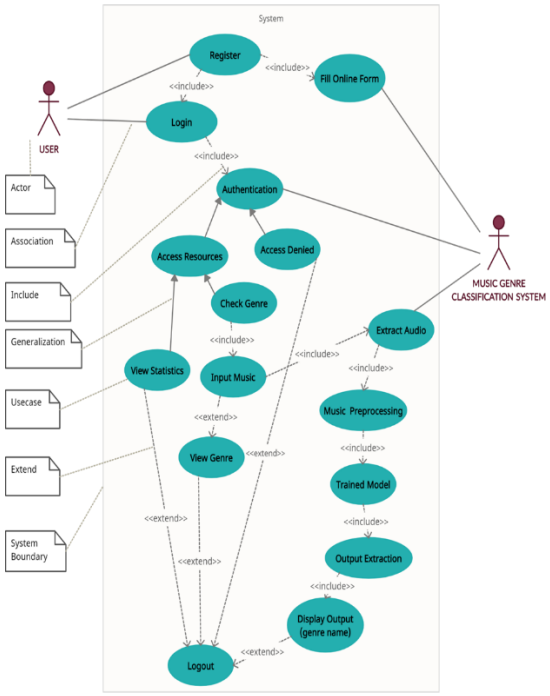


Figure 2 Block Diagram (Music Genre Classification System)

The music genre classification system takes the role of authenticating the users. It stores all the details collected from the users during registration and validates the credentials entered by the user during login. The system then extracts the music file uploaded by the user. It then preprocesses the data and performs feature extraction. The system then identifies the genre of the music based on the features collected and knowledge from the trained model. The system then computes the accuracy of the classification and displays the genre of music and the statistics.

UML USECASE DIAGRAM



This acts as the master of this website which provides certain services. This maintains the details of all the user who have registered to the website. This system then validates the users based on the details collected during the registration. This system takes the music files provided by the user to perform the classification task. The system performs music preprocessing followed by feature extraction. It then classifies the music by comparing the features extracted with the trained model and produces the result. The system then displays the genre of music and the statistics, the total files, training set and testing set count and the accuracy obtained.

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

USER INTERFACE DESIGN

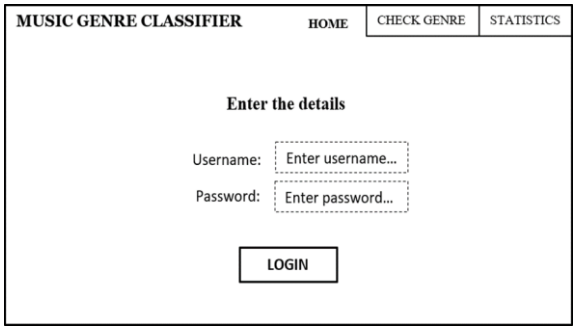


Figure 4 User Interface (Home page)

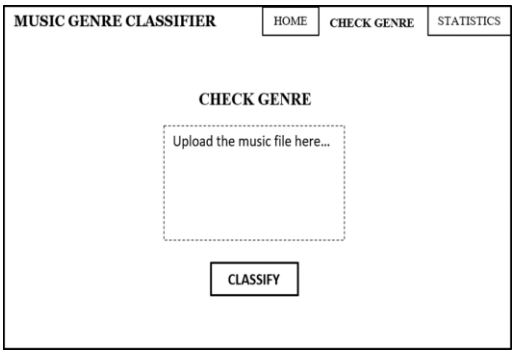


Figure 5 User Interface (Check Genre page)

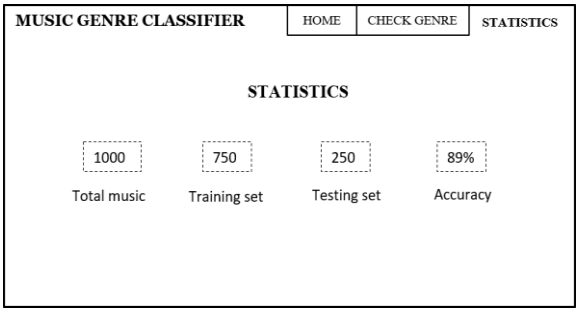


Figure 6 User Interface (Statistics page)

4. Proposed System

DATASET COLLECTION

The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format. The genres are: Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock

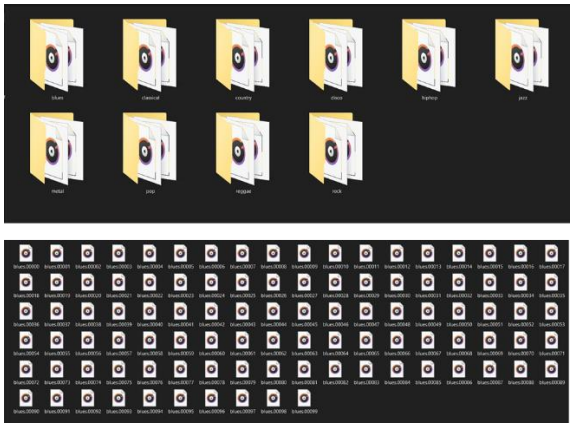


Figure 7 GTZAN Dataset

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

These audio files must be converted into a form which can be interpreted by humans as well as the system. This involves the conversion of audio files into Mathematical Numbers which results in easier interpretation. This can be done by using the librosa package in python. Librosa is an open-source python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.

DATASET VISUALIZATION

The audio file can be interpreted in the form of an image for better visualization. There are two ways to visualize an audio file. They are:

1. Waveform:

The generic term waveform means a graphical representation of the shape and form of a signal moving in a gaseous, liquid, or solid medium. For sound, the term describes a depiction of the pattern of sound pressure variation (or amplitude) in the time domain. In colloquial speech, waveform audio is often used to mean the recorded sound itself (not the graphical representation) in order to distinguish it from structured audio, e.g., MIDI (Musical Instrument Digital Interface) data. The temporal frequencies of sound waves are generally expressed in terms of cycles (or kilocycles) per second. The simplest waveform is the sine wave, since it has only one frequency associated with it. The sound waves associated with, say, music, are constantly varying.

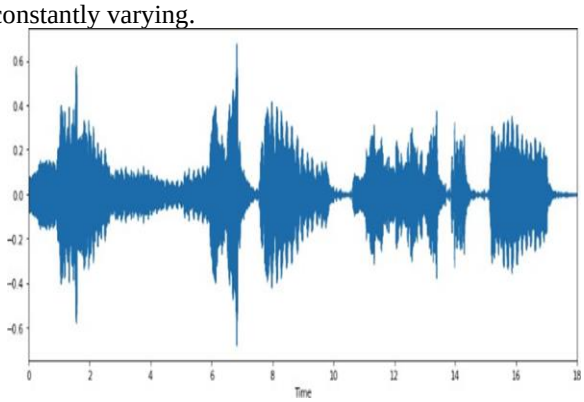


Figure 8 Audio Waveform

2. Spectrogram:

A spectrogram is a visual representation of the spectrum of frequencies of sound or other signals as they vary with time. The Mel spectrogram can be thought of as a visual representation of an audio signal. Specifically, it represents how the spectrum of frequencies vary over time. The Fourier transform is a mathematical formula that allows us to convert an audio signal into the frequency domain. It gives the amplitude at each frequency, and we call this the spectrum. Since frequency content typically varies over time, we perform the Fourier transform on overlapping windowed segments of the signal to get a visual of the spectrum of frequencies over time. This is called the spectrogram. Finally, since humans do not perceive frequency on a linear scale, we map the frequencies to the mel scale (a measure of pitch), which makes it so that equal distances in pitch sound equally distant to the human ear. What we get is the mel spectrogram.

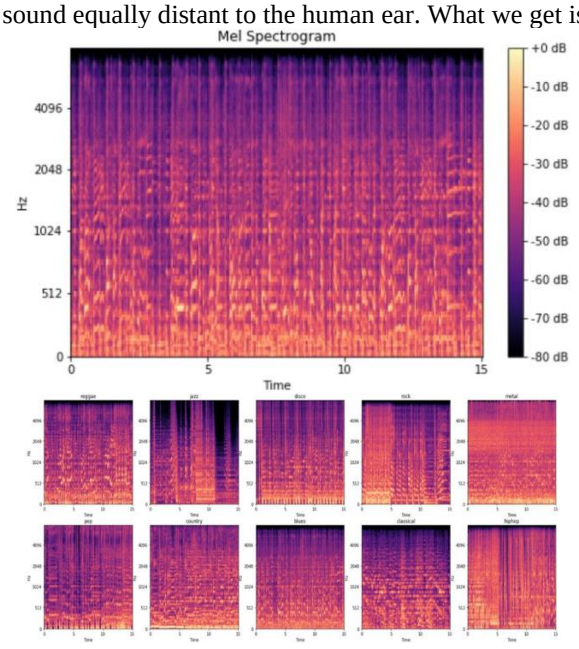


Figure 9 Mel Spectrogram

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

FEATURE EXTRACTION

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. In the GTZAN dataset, a total of 9 features can be extracted. They are:

1. Tempo

Tempo refers to the speed of a musical piece. More precisely, tempo refers to the rate of the musical beat and is given by the reciprocal of the beat period. Tempo is often defined in units of beats per minute (BPM). In classical music, common tempo markings include grave, largo, lento, adagio, andante, moderato, allegro, vivace, and presto. Tempo can vary locally within a piece. Therefore, we introduce the tempogram as a feature matrix which indicates the prevalence of certain tempi at each moment in time.

2. Beats

The beat is the basic metric level in music. It corresponds to the rate at which most people would tap their foot on the floor while listening to music. Beat times correspond to the points in time when the foot would hit the floor.

3. Chroma STFT

Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency (i.e. the original octave) can give useful musical information about the audio -- and may even reveal perceived musical similarity that is not apparent in the original spectra.

4. Root Mean Square (RMS) Energy:

The energy of a signal corresponds to the total magnitude of the signal. For audio signals, that roughly corresponds to how loud the signal is. The energy in a signal is defined as

$$\sum_n |x^2(n)|$$

The root-mean-square energy (RMSE) in a signal is defined as

$$\sqrt{\frac{1}{N}\sum_n |x^2(n)|}$$

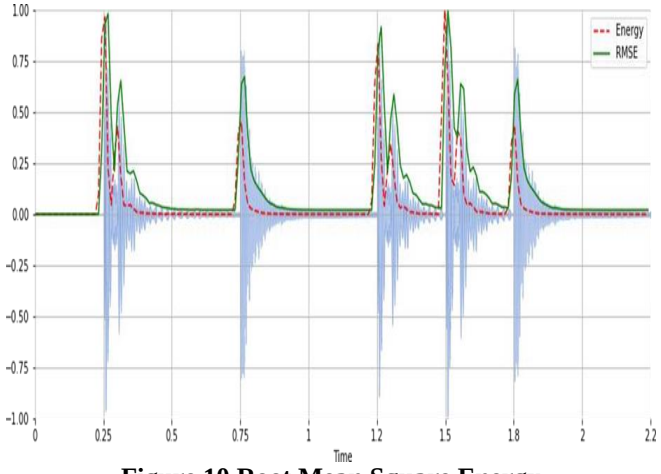


Figure 10 Root Mean Square Energy

5. Spectral Centroid:

It indicates where the" center of mass" for a sound is located and is calculated as the weighted mean of the frequencies present in the sound. Consider two songs, one from a blues genre and the other belonging to metal. Now as compared to the blues genre song which is the same throughout its length, the metal song has more frequencies towards the end. So spectral centroid for blues song will lie somewhere near the middle of its spectrum while that for a metal song would be towards its end. The formula used to calculate centroid is:

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

$$S_{centroid} = \frac{\sum_{K=0}^{\frac{K}{2}-1} K\, A(K)}{\sum_{K=0}^{\frac{K}{2}} A(K)}$$

There is a rise in the spectral centroid towards the beginning.

6. Spectral Bandwidth:

The spectral bandwidth is defined as the extent of the power transfer function around the center frequency. The formula used to calculate spectral bandwidth is:

$$\left(\sum_{k} S(k)\,(f(k) - f_c)^p\right)^{\frac{1}{p}}$$

where S(k) is the spectral magnitude at frequency bin k , f(k) is the frequency at bin k , and fc is the spectral centroid. When p=2 , this is like a weighted standard deviation.

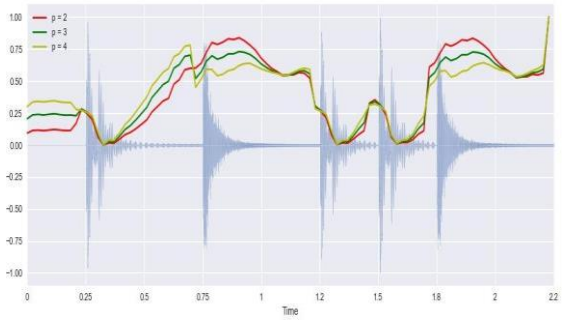


Figure 11 Spectral Bandwidth

7. **Spectral Roll - off:**

Spectral roll - off is the frequency below which a specified percentage of the total spectral energy, e.g. 85%, lies.

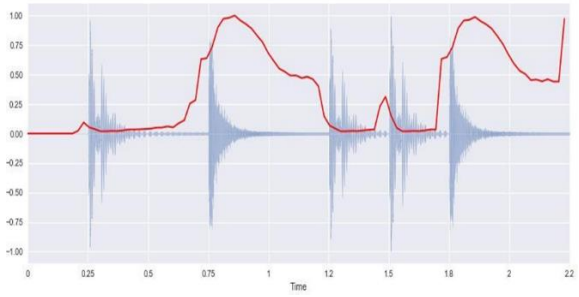


Figure 12 Spectral Roll-off

8. Zero Crossing Rate:

The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval. It usually has higher values for highly percussive sounds like those in metal and rock.

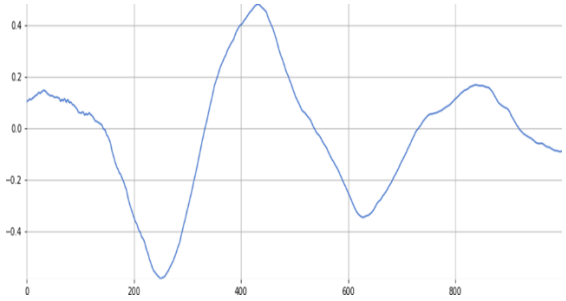


Figure 13 Zero Crossing Rate

This indicates that it has 5 zero crossings.

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

9. Mel-Frequency Cepstral Coefficients:
The Mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice.

$$x(n) \xrightarrow{FFT} A(k) \xrightarrow{\text{Mel band}} M(k) \xrightarrow{log} \log M(k) \xrightarrow{DCT} MFCC(n)$$

The conversion from linear frequency values f to Mel values m is given by a logarithmic function:

$$m = 2595\log_{10}(1 + f/700)$$

The spectral values on the linear frequency scale are integrated in triangular windows which are uniformly spaced on the Mel scale (i.e., they are logarithmically spaced on the linear frequency scale):

$$M(k') = \sum_{k=0}^{k/2-1} A(k) w_{k'}(k)$$

where $w_{k'}(k)$ are triangular windows with increasing width for higher k. We computed 20 MFCC s over 97 frames. The final dataset with the extracted features is:

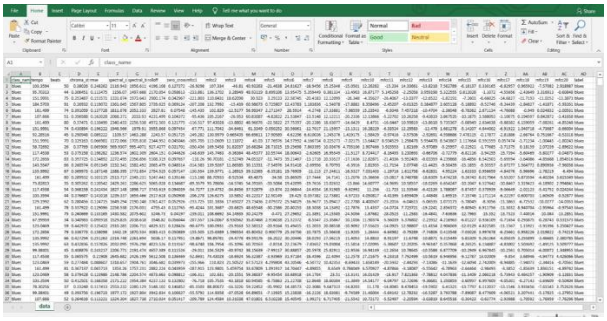


Figure 14 Final Dataset with Extracted Features

5. Experimental results
On surveying about 20 papers on Music Genre Classification, we came to a conclusion that Support Vector Machine algorithm performed better on the dataset than other algorithms and was computationally efficient to implement. We also decided to implement a few algorithms like K-Nearest Neighbor, Random Forest Classifier, Decision Tree Classifier, XGBoost Classifier, Gradient Boost Classifier, Naïve Bayes Classifier and Logistic Regression.

1 RANDOM FOREST:
Random forests is a learning algorithm that is supervised. It's suitable for both classification and regression. Trees make up a forest. A forest is said to be more durable the more trees it contains. Random forests construct decision trees from randomly chosen data samples, get predictions from each tree, and vote on the best solution. It also serves as a useful indicator of the feature importance.
The parameters that provided the best results are:
Criterion -> entropy
max_depth -> 11
max_features -> auto
n_estimators -> 1000
The accuracy obtained is 65.2 %.

2 K-NEAREST NEIGHBOR:
KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure determined from the dataset. This will be very helpful in practice where most of the real-world datasets do not follow mathematical theoretical assumptions. Lazy algorithm means it does not need any training data points for model generation. All training data used in the testing phase. This makes training faster and testing phase slower and costlier. Costly testing phase means time and memory. In the worst case, KNN needs more time to scan all data points and scanning all data points will require more memory for storing training data.
The parameters that provided the best results are:
metric -> manhattan
n_neighbors -> 5
weights -> distance
The accuracy obtained is 62 %.

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

3 SUPPORT VECTOR MACHINE:
Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyper plane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.
The parameters that provided the best results are:
C -> 100
degree -> 1
gamma -> 1
kernel -> rbf
The accuracy obtained is 66.4 %.

4 DECISION TREE:
Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.
The parameters that provided the best results are:
criterion -> gini
max_depth -> 9
min_samples_leaf -> 4
The accuracy obtained is 49.2 %.

5 XGBOOST:
XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Basically, XGBoost is an algorithm. Also, it has recently been dominating applied machine learning. XGBoost is an implementation of gradient boosted decision trees. Although, it was designed for speed and performance. Basically, it is a type of software library. That you can download and install on your machine. Then have to access it from a variety of interfaces.
The parameters that provided the best results are:
booster -> gbtree
eval_metric -> rmse
objective -> multi:softmax
The accuracy obtained is 64.8 %.

6 GRADIENT BOOST:
Gradient boosting re-defines boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimizing a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc. Intuitively, gradient boosting is a stage-wise additive model that generates learners during the learning process (i.e., trees are added one at a time, and existing trees in the model are not changed). The contribution of the weak learner to the ensemble is based on the gradient descent optimisation process. The calculated contribution of each tree is based on minimising the overall error of the strong learner.
The parameters that provided the best results are:
N_estimators -> 25
The accuracy obtained is 60 %.

7 NAÏVE BAYES:
Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.
P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.
P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.
P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability
The accuracy obtained is 40.8 %.

8 LOGISTIC REGRESSION:
Multinomial logistic regression is an extension of logistic regression that adds native support for multi-class classification problems. Logistic regression, by default, is limited to two-class classification problems. Some extensions like one- vs-rest can allow logistic regression to be used for multi-class classification problems, although they require that the classification problem first be transformed into multiple binary classification problems. Instead, the multinomial logistic regression algorithm is an extension to the logistic regression model that involves changing the loss function to cross-entropy loss and predict probability distribution to a multinomial probability distribution to natively support multi-class classification problems.
The parameters that provided the best results are:
C -> 11.288378916846883
Penalty -> l2
Solver -> lblinear
The accuracy obtained is 60 %.
The Accuracy of the above-mentioned algorithms is shown below:

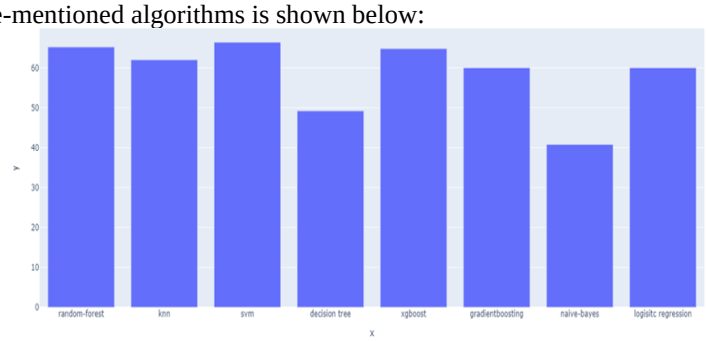


Figure 15 Accuracy Obtained by Machine Learning Algorithms

SUPPORT VECTOR MACHINE:
Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

WORKING OF SUPPORT VECTOR MACHINE:
The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset.SVM searches for the maximum marginal hyperplane in the following steps:
1. Generate hyperplanes which segregates the classes in the best way.
2. Select the right hyperplane with the maximum segregation from the either nearest data points.

HYPERPARAMETERS:
i. Kernel: The main function of the kernel is to transform the given dataset input data into the required form. There are various types of functions such as linear, polynomial, and radial basis function (RBF). Polynomial and RBF are useful for non-linear hyperplane. Polynomial and RBF kernels compute the separation line in the higher dimension. In some of the applications, it is suggested to use a more complex kernel to separate the classes that are curved or nonlinear. This transformation can lead to more accurate classifiers.
ii. Regularization: Regularization parameter in python's Scikit-learn C parameter used to maintain regularization. Here C is the penalty parameter, which represents misclassification or error term. The misclassification or error term tells the SVM optimization how much error is bearable. This is how you can control the trade-off between decision boundary and misclassification term. A smaller value of C creates a small-margin hyperplane and a larger value of C creates a larger-margin hyperplane.

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

iii. Gamma: A lower value of Gamma will loosely fit the training dataset, whereas a higher value of gamma will exactly fit the training dataset, which causes over-fitting. In other words, you can say a low value of gamma considers only nearby points in calculating the separation line, while the value of gamma considers all the data points in the calculation of the separation line.

The audio file to be classified is obtained from the user through the check genre web-page build. The user will be able to upload any file into the webpage. It is the role of the webpage to check if it is a valid audio file or not.

i. If the file is not of the .wav format, an error pops up as a toast message stating that the file is invalid.
ii. If the user uploads, a valid .wav file, the file gets uploaded and commences the classification process.

The audio file to be classified is extracted from the user interface where the user has uploaded the file. The features are extracted from the audio file and the same process for creation of dataset sticks to this as well. The model created is then loaded and the features thus generated are used to classify the audio file to its proper genre.
The final result is the genre of the audio file uploaded by the user. This is displayed in the check genre page of the website as an alert dialog or a toast message. The final result also includes the statistics of the dataset count, testing and training count and the accuracy of the implemented system.

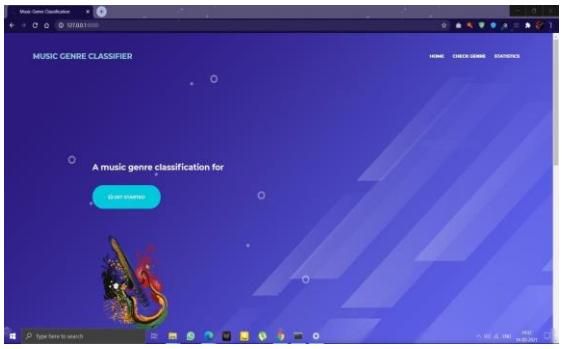


Figure 16 Music Genre Classification (home page)

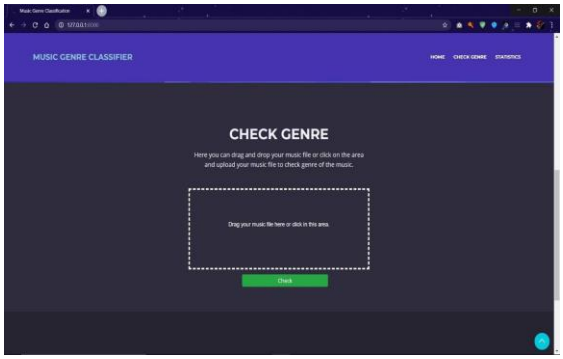


Figure 17 Music Genre Classification (check genre page)

CASE 1: when user uploads an invalid file:
In this case, the system doesn't accept the file and notifies the user with a toast message stating 'only .wav file type is allowed'.

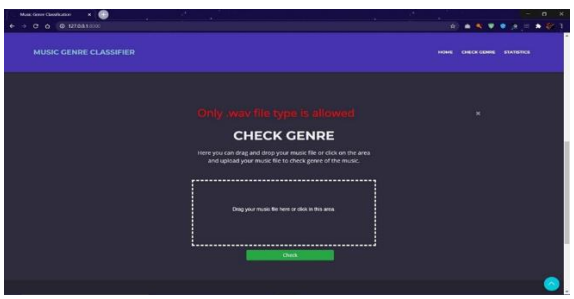


Figure18 Uploading an Invalid File

CASE 2: when user uploads a valid file:
In this case, the system extracts the .wav audio file and proceeds on the classification process.

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

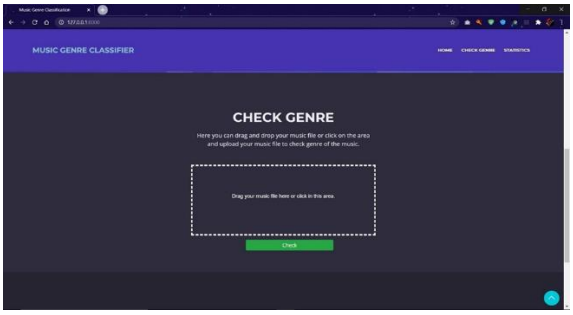


Figure 19 Uploading a Valid Audio File

CLASSIFICATION OF MUSIC GENRE

Once the user uploads a valid file, the system commences the classification process. The features are extracted from the test file and the classification model classifies the audio file into the correct genre.

VIEW RESULTS

Once the music is classified by the system, the music genre of the audio file is displayed to the user as the result.

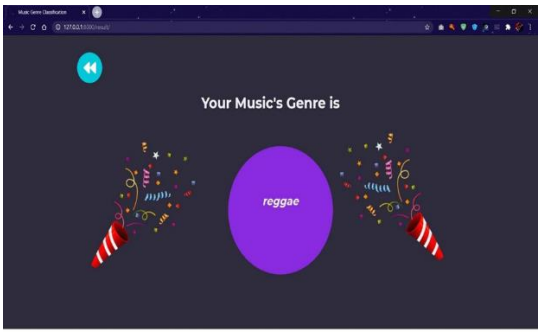


Figure 20 View Results

VIEW STATISTICS

The system also displays the details of the following: total count of files in the dataset, the training and testing split up, the accuracy obtained by the system.

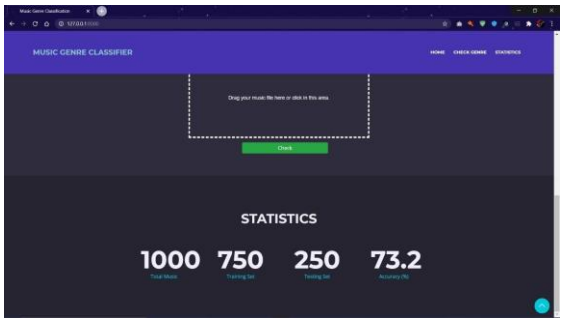


Figure 21View Statistics

CONCLUSION

The exponential growth in the internet and multimedia systems applications that categorizes music based on genre has led us to develop a system for the above task. Our system can help apps which organize and classify songs, albums, and artists into broader groups which shares similar musical characteristics. Automatic analysis and classification of the music is one of the required components of such Music information retrieval systems. The proposed system to classify music reduces laborious manual work and automates the task of classifying music. Hence, we have proposed system covering the stated functionality with optimal accuracy. Various techniques have been employed in each phase in this system to classify the music. Challenges still prevail in the system like classifying a music genre which is new to the system and input file with more noise. The proposed system has shown enhanced performance in classifying music genre.

ISSN-2394-5125 VOL 07, ISSUE 17, 2020

REFERENCES

[1] Wu, M. and Liu, X., 2020, January. A Double Weighted KNN Algorithm and Its Application in the Music Genre Classification. In 2019 6th International Conference on Dependable Systems and Their Applications (DSA) (pp. 335-340). IEEE.

[2] Liang, B. and Gu, M., 2020, August. Music genre classification using transfer learning. In 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 392-393). IEEE.

[3] Ghosal, S.S. and Sarkar, I., 2020. Novel Approach to Music Genre Classification using Clustering Augmented Learning Method (CALM). In AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1).

[4] Vishnupriya, S. and Meenakshi, K., 2018, January. Automatic Music Genre Classification using Convolution Neural Network. In 2018 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-4). IEEE.

[5] Jawaherlalnehru, G. and Jothilakshmi, S., 2018. Music Genre Classification using Deep Neural Networks. International Journal of Scientific Research in Science, Engineering and Technology, 4(4), p.935.

[6] Quinto, R.J.M., Atienza, R.O. and Tiglao, N.M.C., 2017, November. Jazz music sub- genre classification using deep learning. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 3111-3116). IEEE.

[7] Zhang, W., Lei, W., Xu, X. and Xing, X., 2016, September. Improved Music Genre Classification with Convolutional Neural Networks. In Interspeech (pp. 3304-3308).

[8] Jeong, I.Y. and Lee, K., 2016, August. Learning Temporal Features Using a Deep Neural Network and its Application to Music Genre Classification. In Ismir (pp. 434-440).

[9] Rajanna, A.R., Aryafar, K., Shokoufandeh, A. and Ptucha, R., 2015, December. Deep neural networks: A case study for music genre classification. In 2015 IEEE 14th international conference on machine learning and applications (ICMLA) (pp. 655-660). IEEE.

[10] Haggblade, M., Hong, Y. and Kao, K., 2011. Music genre classification. Department of Computer Science, Stanford University, 131, p.132.

[11] Yaslan, Y. and Cataltepe, Z., 2006, August. Audio music genre classification using different classifiers and feature selection methods. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 2, pp. 573-576). IEEE.

[12] Koerich, A.L. and Poitevin, C., 2005, October. Combination of homogeneous classifiers for musical genre classification. In 2005 IEEE International Conference on Systems, Man and Cybernetics (Vol. 1, pp. 554-559). IEEE.

[13] Meng, A., Ahrendt, P. and Larsen, J., 2005, March. Improving music genre classification by short time feature integration. In Proceedings (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 5, pp. v-497). IEEE.

[14] Li, T., Ogihara, M. and Li, Q., 2003, July. A comparative study on content-based music genre classification. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 282-289).

[15] Tzanetakis, G. and Cook, P., 2002. Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5), pp.293-302.

[16] Bahuleyan, H., 2018. Music genre classification using machine learning techniques. arXiv preprint arXiv:1804.01149.

[17] Li, T.L., Chan, A.B. and Chun, A.H., 2010. Automatic musical pattern feature extraction using convolutional neural network. Genre, 10, p.1x1.

[18] Paradzinets, A., Harb, H. and Chen, L., 2009. Multiexpert system for automatic music genre classification. Teknik Rapor, Ecole Centrale de Lyon, Departement MathInfo.

[19] Jang, D., Jin, M. and Yoo, C.D., 2008, June. Music genre classification using novel features and a weighted voting method. In 2008 IEEE International Conference on Multimedia and Expo (pp. 1377-1380). IEEE.

[20] Lu, L., Zhang, H.J. and Jiang, H., 2002. Content analysis for audio classification and segmentation. IEEE Transactions on speech and audio processing, 10(7), pp.504-516.