Music Genre Prediction with similarity and Instrument Analysis

Shivani Gurung Rakesh
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
shivani.gurung@se21.qmul.ac.uk

***Abstract*—In this paper, I'll examine the audio features extracted from the dataset and build various types of ensemble models to see how well we can distinguish one genre from another and differentiate instruments. Used two data sets GTZAN dataset and the IRMAS dataset to apply these approaches and features to this study. Investigating music sound documents because of types and other subjective labels is a functioning field of exploration in AI. When matched with individual order calculations, most quite support vector machines (SVMs) and k-closest neighbor classifiers (KNNs), certain highlights, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma attributes, and other spectral properties, have been demonstrated to be powerful elements for grouping music by type. To discover trends and patterns from massive volumes of data, machine learning techniques have shown to be quite successful. Analyzing music also follows the same rules.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Python has some excellent audio processing libraries, such as Librosa and PyAudio. There are also modules built in for some basic audio functions. In this paper, I'd like to look at some of the fundamental differences between various musical genres. Experts have been attempting to comprehend sound and what distinguishes one from another.

The purpose of this study is to present the performance evaluation of various classification methods for Music Genre classification using a supervised learning approach. As this is a multi-class classification category, the correlation between classes is important in the classification task.

Music can be classified into a variety of genres,but what is genre? A genre is a classification system that differentiates music into styles based on their most identifiable qualities. The music or song in the same genre have similar forms and styles. Genres are broken into sub-genres For more precise categorization, however the fundamental purpose is to effectively categorise it into its designated category.

Although no classifier can correctly classify all music samples based on their various features, we attempt to demonstrate the impact of training these classifiers on various combinations of music datasets. Music Genre Classification is a well-known machine learning problem with numerous applications. It can be used to tag every song in a massive music corpus with genre and sub-genre, which can then be used to find similar songs. A music recommendation system is another application. We can cluster similar songs and share them with users based on their preferences by using features from an intermediate layer in the model.

The advantages of this classification method include ease of implementation, ease of use, and speedy retrieval, but there are also clear disadvantages. First off, this approach relies on manually labeled music data, which is labor-intensive and difficult to do correctly without labeling music information issues. Second, the music's audio data is not used in this text-based approach. Pitch, timbre, and melody are just a few of the essential aspects of music that are included in audio data. The data comprises songs from several decades of the last century, resulting in a broad range of audio quality.

II. Related Work

Proper music classification is critical for increasing the effectiveness of music information retrieval. At the moment, music classification primarily consists of text classification and music content classification. Text classification is primarily based on music metadata such as singer, lyrics, songwriter, age, music name, and other labelled text information. [12]

A comparable situation of music retrieval is audio alignment, matching, or synchronisation, in which the goal is to locally connect temporal locations from two music signals in addition to recognising a specific audio fragment. Furthermore, depending on the strength of the audio qualities, multiple performances of the same piece might be aligned. [6]

The method used here in this journal [7] to identify the instrument involved extracting characteristics from the melspectrogram using CNN convolutional layers. Mel spectrograms are visual representations of musical spectrum qualities such as playing technique, sound excerpt frequency, and other spectral features.

The magnitude of the natural logarithm-compressed mel-frequency spectrogram is the input supplied to the CNN. The majority of the information from a sound snippet is extracted using a variety of sampling techniques and modifications, which can be referred to in this work. Convolutional layer that automatically extracts features from spectrograms and max pooling are both employed in the proposed CNN architecture, which is intended to detect instruments.'ReLU' (alpha = 0.33) provided the best results with several activation functions. On the dataset IRMAS, the classification result had an overall

F score of 0.60. [8] They investigated how five distinct woodwind instruments were categorised. Mel frequency cepstral coefficient (MFCC) characteristics were retrieved from the training tracks since they were found to be beneficial for tremolo, vibrato, and sound attack categorization. PCA was used to reduce dimension on the MFCC features before feeding the modified features into (GMM) and (SVM) classification. SVM was also evaluated using linear and polynomial kernels, with the latter showing to be more efficient.

III. Methodology

A. Understanding Audio Files

Two different datasets have been used for training and testing. For genre prediction GTZAN dataset and for instrument analysis IRMAS dataset is used.

- IRMAS - IRMAS is designed to help train and test methods for automatically recognizing the predominant instruments in musical audio. This data is polyphonic, allowing for the construction of a strong classifier. The data consists of roughly eleven different instruments in.wav files with a length of three seconds each. Six of these instruments have been selected for acclaim. The 3846 music samples in this collection, which span around three hours, provide enough information for both training and testing purposes. The data also includes a variety of musical styles, such as pop-rock, classical, country folk, and latin soul. The inclusion of these many genres could improve training. This dataset comes by Ferdinand Fuhrmann in his PhD thesis (IRMAS: a dataset for instrument recognition in musical audio signals - MTG - Music Technology Group (UPF), 2022) The dataset consists audio of instruments like acoustic guitar, electric guitar, organ, piano,cello, clarinet, flute,saxophone, trumpet, violin, and human singing voice. [1] For this project The data consists of roughly eleven different instruments in .wav files with a duration of three seconds each. Six of these instruments have been selected for analysis (flute, piano, trumpet, guitar,voice, organ).
- GTZAN - The gtzan audio dataset incorporates a thousand tracks of 30 second length. There are 10 genres, every containing a hundred tracks that are all 22050Hz Mono 16-bit audio files in .wav format. It consists of approximately 100 tracks per genre from a collection of 1000 tracks of ten different genres. It consists of different genres like blue, classical, country, disco, hip-hop,jazz, metal, pop, reggae, and rock. One folder in the data set contains for each song (30 seconds long) the mean and variance computed over multiple features extracted from audio files. The other folder contains two CSV files containing the audio files.
  The second file has the same structure as the previous one, but the songs were originally divided into 3 second audio files (this multiplied the amount of data we have to work with). The more data we have, the better our classification model will perform. [**?**], [**?**]

B. Data Preprocessing

Data preprocessing is an essential process that improves the data quality to support the extraction of meaningful information the data's insights It speaks of the method of preparation. preparing the raw data for use by (cleaning and organising) creating and educating machine learning models Simply put Data preprocessing, in other words it is data mining technique that converts unprocessed data into a readable and understandable format. Working with audio files we need to basic understanding about characteristics of audio signal.The characteristics of an audio signal include its bandwidth, frequency, decibel level, etc. The amplitude and time functions can be used to express a typical audio signal.

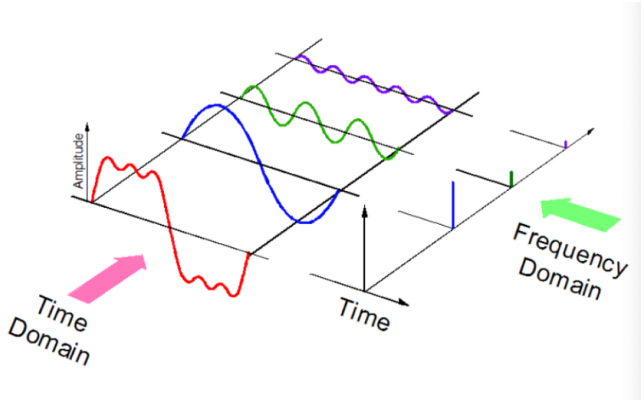

Fig. 1. Audio Processing with Python
[9]

C. feature extraction

The audio signals are characterised by their features. To train the machine learning model, we have used the files containing the features of the audio clips. These features are extracted using the librosa, a library of python. The audio features are broadly classified into two categories, Time domain and Frequency domain.

The Python library here is used for analysing audio and musical data is called Librosa. Load, Display, and Features are a few of librosa's crucial features. The "Load" command loads an audio file as a time series in floating point. Utilizing Matplotlib, "Display" offers visualisations including waveform and spectrogram. MFCC and other spectral features are extracted and modified using the 'Features' command. By converting from frequency (Hz) scale to mel-scale, MFCCs are produced. For achieving more accurate classification of musical instruments, it is necessary to extract more complicated features apart from MFCC. Hence considered other features like Zero-crossing rate, Spectral centroid, Spectral bandwidth, and Spectral mode.

it is necessary to extract more intricate features in addition to MFCC in order to classify musical instruments more accurately. As a result, we took into account additional features throughout our feature extraction process using Li-

brosa, including Zero-crossing rate, Spectral centroid, Spectral bandwidth, and Spectral roll-off [3].

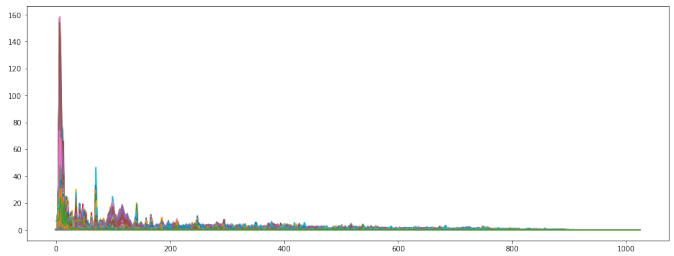

Fig. 2. Fourier Transformation

Fourier Transform - It divides a signal into discrete spectral components and offers frequency information about the signal as a result. Then transforms the amplitude as well as the frequency both y-axis to Decibels, which approximates the log scale of amplitude.

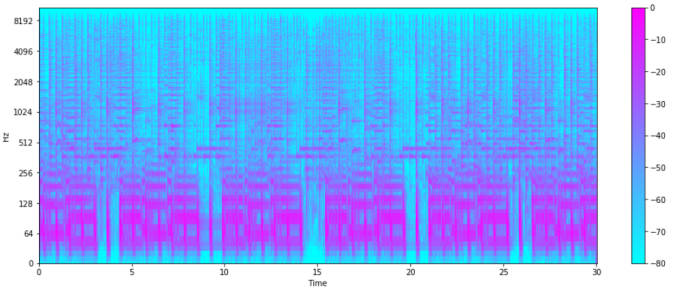

Fig. 3. spectogram

An illustration of a signal's frequency spectrum as it changes over time could also be called as sonographs.

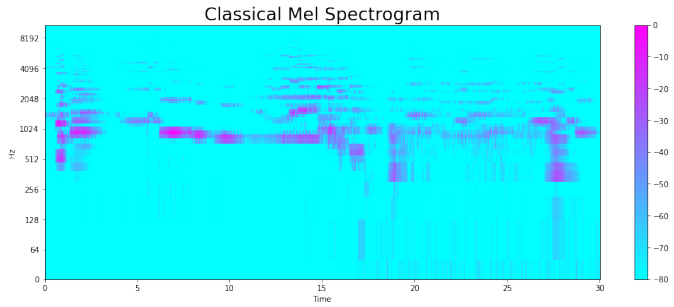

Fig. 4. classical mel spectogram

The rate at which the signal crosses zero is indicated by the zero crossing rate. Spectral Centroid is a metric used to determine the centre of mass of a spectrum, based on the appearance of brightness in a sample. [8] The weighted average of the frequency signal by its spectrum is given by spectral bandwidth. The frequency to which a certain proportion of the total spectral energy belongs is referred to as spectral roll-off.

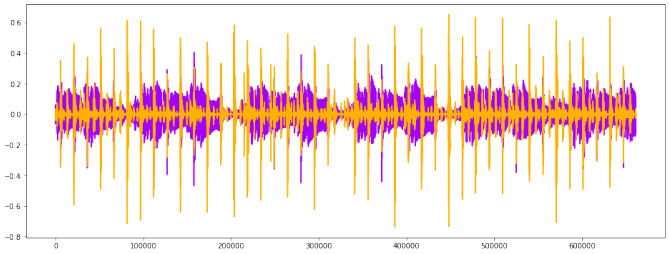

Fig. 5. harmonics and perceptual

An escalating succession of acoustic elements of the audio that are perceptible above the fundamental frequency.

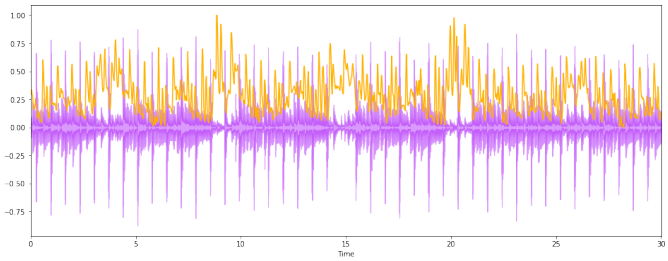

Fig. 6. spectral centroid with wave form

The Fourier transform frequency and amplitude information is used to compute the amplitude at the centre of the spectrum of the signal distribution over a window.

The Librosa package was among the very first Python packages developed to analyze audio data. It is more popular as well as longer established and has offered greater out-of-sample validation accuracy. [4]To study and visualise the audio in this project, librosa will be used. Used spectogram and waveform to visualise the audio files. The frame size and hopsize (hop length between frames) default inputs are used to extract MFCCs from the audio spectrum for analysis. The default values for the sampling rate, hopSize, and frame size in Essentia are 44.1 kHz, 512, and 1024 respectively. The characteristics that have been successfully retrieved from various signal sample signal segments are then averaged. Once each sample has been identified with its instrument class, they are utilised as its features.

One of the crucial steps in the whole process is to extract the right features that can help us in distinguishing all these genres with the best prediction rate. Different feature extraction methods [9] were used to extract the features based on different audio features. In order to improve the prediction rate, we use different feature extraction methods based on the characteristics of the different audio features in order to extract the features that can help us distinguish all the genres correctly.

Sound is represented as an audio signal with parameters such as frequency, bandwidth, decibel, and so on. Amplitude and Time can be used to express a typical audio signal.

now lets explore the csv file of the data set GTZAN:

the CSV file basically consists of Extracted audio features such as MFCC(Mel-Frequency Cepstral Coefficients), Zero

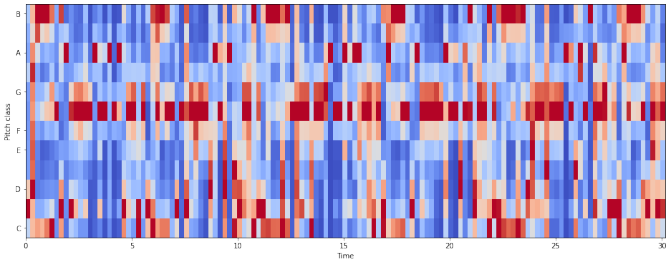

Fig. 7. Chroma frequencies

The chroma representation provides information on the intensity of each of the 12 different musical chroma at each instant in time.

Crossing Rate, Harmonics and Perceptrual,Spectral rolloff and Chroma frequencies.

D. Exploratory Data Analysis (EDA)

is a way of evaluating datasets to summarise their essential properties, frequently using visual approaches. EDA is used before modeling in order to visualize the properties of the audio by utilizing summary statistics and graphical representations to [10] conduct preliminary investigations on data in order to discover correlations, spot outliers, test hypotheses, and check assumptions.

The (features30sec).csv in the GTZAN data will be subjected to EDA. This file basically provides the mean and variance for each audio file concerning the above-mentioned attributes of the audio.

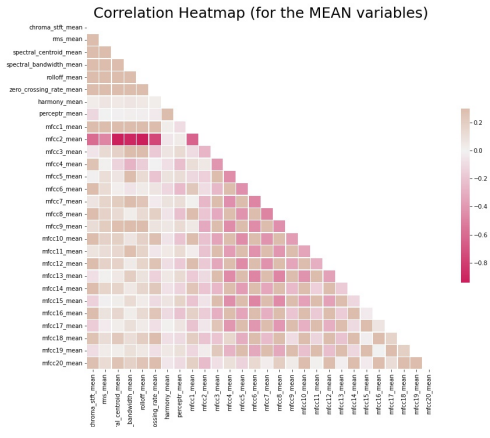

Fig. 8. correlation heat map

Matrix of Correlation Heat map is not just a feature selection technique but rather a feature visualisation tool that may be used to pick and engineer features that is used for training further into Machine Learning algorithms.

Some of the features are as follows:

E. Label Encoding

label Encoder performs the conversion of these labels of categorical data into a numeric format. It refers to converting the labels into a numeric form so as to convert them into the machine readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. In our dataset, we have the column 'label' which contains the genres of the respecive to the audio signals. Thus, these categories need to be encoded into numerical form. This is done with the [11] function LabelEncoder() of sklearn.preprocessing. The labels are encoded as follows:

(JPG to be attached once the code is compiled and output is optained)

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \quad \textbf{(Eq.1)}$$

Fig. 9. Covariance Formula

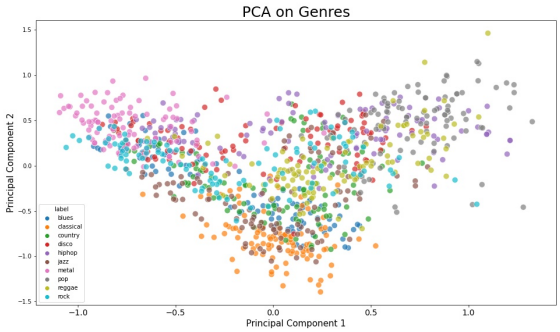

Fig. 10. PCA Statter plot

here we do pca analysis inorder to visualise the genre groups from the GTZAN dataset.

F. Feature Scaling, Train and splitting the dataset

Feature scaling is a technique used to put the independent features of the data into a range of independent variables or features into predetermined range. Standard scaler scales the data so that the distribution is centered around 0, with a standard deviation of 1, and makes the assumption that the data are normally distributed within each feature by removing the mean and scaling [12] to unit variance which is required for machine learning estimators.

Splitting the dataset into two separate sets – training set and test set. This process involves partitioning a dataset into two subsets. The first subset, known as the training dataset, is utilized to fit the model. The second subset is not used to train the model instead, the model is given the dataset's input element, and predictions are generated and compared to predicted values. The second dataset is known as the test-dataset. The ML model for genre prediction the data used is split into 70% - 30% and for instrument detection 80% - 20% to predict outcomes.

In order to judge the performance of the model, Prescision, Recall and F1 score were used. Precision is the ratio in the figure 12 , where tp(true positive)is the number of true posi-

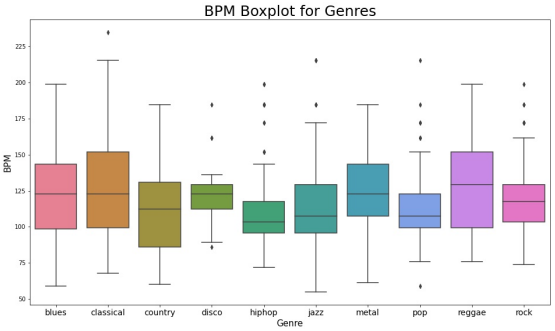

Fig. 11. Box Plot

tives and f p(false positive) the number of false positives.Here the classifier does not label false positive classifier as positive.

$$\frac{tp}{tp + fp}$$

Fig. 12. Precision

$$\frac{tp}{tp + fn}$$

Fig. 13. Recall

Recall is the ratio in the figure 13 where tp (true positive) is the number of true positives and fn(false negative) the number of false negatives. Recall is helps the classifier to find the positive values.

F1 score 14 can also be called as the harmonic mean of Precision and Recall.

$$f1 = \frac{2(precision * recall)}{precision + recall}$$

Fig. 14. F1 Score

IV. RESULT

For the genre prediction tried ten models and trained, XG-booster had high accuracy performance than the rest of the models. this figure 15 here is a confusion matrix output.

As for the recommender system to determine the best similarity for each given vector, sorted from the best match to the least good match. Has been accomplished for audio files using the cosine similarity library. in fig 16 For each combination of songs in the data, get the pairwise cosine similarity. Then takes a song's name as input and returns the top 5 songs as an output that are most similar to it. For the instrument analysis after comparing SVM, Random Forest, and XG-Boost

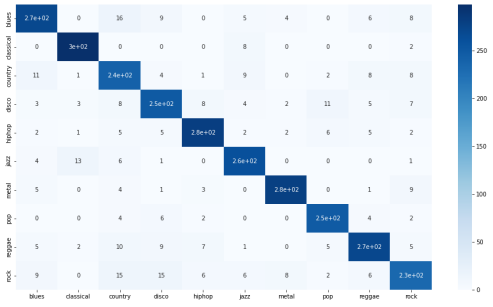

Fig. 15. Confusion Matrix of XG-Boost classifier

$$Similarity(A, B) = \cos(\theta) = \frac{A \cdot B}{||A|| \times ||B||} = \frac{18}{\sqrt{17} \times \sqrt{20}} \approx 0.976$$

Fig. 16. similarity formula used

classifier the best performing classifier is the SVM classifier with the accuracy score of 78% whereas in the XG-boost has a accuracy of 69% 19 and random forest classifier has accuracy of 18 the accuracy is 67% comparatively low in comparision. in the figure 17, 18 and 19 i have obtained a confusion matrix for the supervised models used.

Fig. 17. SVM
SVM with accuracy of 78%

REFERENCES

[1] www.upf.edu. (n.d.). IRMAS: a dataset for instrument recognition in musical audio signals - MTG - Music Technology Group (UPF). [online] Available at: https://www.upf.edu/web/mtg/irmas.

[2] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), pp.293–302. doi:10.1109/tsa.2002.800560.

[3] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), 2000, pp. II753-II756 vol.2, doi: 10.1109/ICASSP.2000.859069.

Fig. 18. RF
Random Forest with accuracy of 70%

Fig. 19. XG-boost
XG-Boost with accuracy of 70%

[4] McFee, B., Kim, J.W., Cartwright, M., Salamon, J., Bittner, R.M. and Bello, J.P. (2019). Open-Source Practices for Music Signal Processing Research: Recommendations for Transparent, Sustainable, and Reproducible Audio Research. IEEE Signal Processing Magazine, 36(1), pp.128–137. doi:10.1109/msp.2018.2875349.

[5] Zalkow, F., Balke, S., Arifi-Müller, V. and Müller, M. (2020). MTD: A Multimodal Dataset of Musical Themes for MIR Research. Transactions of the International Society for Music Information Retrieval, 3(1), pp.180–192. doi:10.5334/tismir.68.

[6] Dixon, S. and Widmer, G. (2005). A music alignment tool chest. In International Conference on Music Information Retrieval.

[7] Han, Y., Kim, J. and Lee, K. (2017). Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1), pp.208–221. doi:10.1109/taslp.2016.2632307.

[8] Essid, S., Richard, G. and David, B. (2006). Musical instrument recognition by pairwise classification strategies. IEEE Transactions on Audio, Speech and Language Processing, 14(4), pp.1401–1412. doi:10.1109/tsa.2005.860842.

[9] Parul Pandey (2018). Music Genre Classification with Python. [online] Medium. Available at: https://towardsdatascience.com/music-genre-classification-with-python-c714d032f0d8.

[10] (Chathuranga and Jayaratne, 2013) Chathuranga, D. and Jayaratne, L. (2013). Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches. GSTF Journal on Computing (JoC), 3(2). doi:10.7603/s40601-013-0014-0.

[11] Nieto, O., Mysore, G.J., Wang, C., Smith, J.B.L., Schlüter, J., Grill, T. and McFee, B. (2020). Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications. Transactions of the International Society for Music Information Retrieval, 3(1), pp.246–263. doi:10.5334/tismir.54.

[12] Kostek, B. (2004). Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques. Proceedings of the IEEE, 92(4), pp.712–729. doi:10.1109/jproc.2004.825903.

MSc Project - Reflective Essay

Project Title:	Music Genre Prediction with similarity and Instrument Analysis
Student Name:	Shivani Gurung Rakesh
Student Number:	210268155
Supervisor Name:	Marcus Pearce
Programme of Study:	ECS750P

The creation of functionalities demands knowledge and in-depth understanding of the audio and music domain. The characteristics of various categorization tasks are sometimes not universal and complete. Because of manual and traditional machine learning approaches, extracting music characteristics is tough when we don't have proper knowledge about the relevant field.

As a result, the contribution of this study is to transform a music audio signal into a proper visualization of a sound spectrum as a so removing the issue of human feature selection for the better understanding of the audio or music dataset. According to a research people usually like recommendations that take music genres into account like pop, rock etc., over those that are merely focused on similarity. Exploring more of MFCC characteristics for the audio data set to train better and attain a higher accuracy score.

The Strength and Weakness of the Project

This classification system has advantages including simple implementation, usability, and quick retrieval, but there are also definite drawbacks. First off, this strategy depends on labour-intensive and manually classified music data Correct execution is challenging without labels for music metadata problems. Second, this text does not make use of the auditory data from the music. based strategy Melody, timbre, and pitch are a few examples of the fundamental components of music that are contained in audio files. The songs from various decades of the previous century are included in the data. as a result, there is a wide variety in audio quality.

SVM

Instrument	P	R	F1
flute	0.60	0.71	0.65
piano	0.81	0.80	0.80
trumpet	0.84	0.76	0.80
guitar	0.82	0.82	0.82
voice	0.75	0.69	0.72
organ	0.78	0.85	0.81

XG-Boost

Instrument	P	R	F1
flute	0.61	0.51	0.55
piano	0.76	0.73	0.75
trumpet	0.60	0.68	0.64
guitar	0.76	0.70	0.73
voice	0.69	0.61	0.65
organ	0.66	0.78	0.72

Random Forest

Instrument	P	R	F1
flute	0.72	0.39	0.50
piano	0.76	0.75	0.76
trumpet	0.65	0.67	0.66
guitar	0.75	0.74	0.74
voice	0.80	0.61	0.69
organ	0.62	0.88	0.73

The above image is the classification report from the instrument analysis (IRMAS dataset).

It is observed with the classification that the model is not able to differentiate much between piano and guitar in all the three classifiers as the values are close. Even though the best performing classifier was SVM but it still needs improvement.The song modelling improves the performance but also decreases the classifying train. Naturally, it might be argued that all conclusions are still similar because all this research had to deal with the same issues in GTZAN. Although this falsely assumes that all machine learning techniques now in use are impacted by these in the same ways. issues. Since these might be divided between training and Exact copies of testing data will typically unfairly exaggerate the effectiveness of various systems.

As for weakness of the project, the data set which we used for genre prediction GTZAN have few faults as researched by (Sturm, 2013) on his journal. GTZAN data has repetitions, labelling errors, and abnormalities that make any result generated using it difficult to comprehend.

Some future directions as pointed out in the previous section include a more complex classifier and extensive parameter tuning for the classifier and dictionary training. If I would have some more time, I would try to use different data set which has less faults or rectify the issues of the dataset and then train them into the model respectively, would also try to implement CNN.

The music recommender system worked pretty well as the songs which had matched based on similarity were actually similar.

Future Scope of music analysis

A better result for this task means that one could possibly have infinite data (since you can generate synthetic data) to train complex and data hungry machine learning models which can obtain a higher classification rate. In future I will be exploring new methods in-order to learn more about sound, audio signals and different kinds of music.

The creation of functionalities demands knowledge and in-depth understanding of the music domain. The characteristics of various categorization tasks are sometimes not universal and complete. Due to old ML approach, getting music features is difficult. This research basically extracts audio features or signal and transforms into visually understanding audio which can also be called as audio signal representation or visualization. We can utilize it in for exploring other traditional instruments like Indian instrument and Chinese instruments audio data set inorder to extract MFCC features. With a complex neural network models and identifying the instruments and exploring more audio features would be more adventurous.

A better result for this task means that one could possibly have infinite data (since you can generate synthetic data) to train complex and data hungry machine learning models which can obtain a higher classification rate. In future I will be exploring new methods in-order to learn more about sound, audio signals and different kinds of music.

The out put show improvement for further use of already established machine learning concepts of audio musical data or song whereas currently machine learning concentrates on removing traditional techniques with DL (deep learning). Each network should be classified and trained in such a way that it perfectly fits the classifier. In coming days a person can also concentrate on using a complex type of data set using a image format of a video format.

By splitting up the song into small chunks and then getting the features for those chunks and then individually training and classifying We could benefit from the majority of the methods' advantages. Very few features vector for a small chunk of audio makes training fast and smaller piece uniformed as timbrally.

Theory and Practical work

The rise and popularity of online streaming services, which currently put practically all of the world's music at the user's fingertips, has fuelled a recent surge in music recommender systems

(MRSs). Even though modern MRSs greatly assist consumers in locating interesting music in these vast archives, MRS research still faces several obstacles. (Schedl et al., 2018). MRS

In order to satisfy the preferences of the audience researching in this domain becomes more important as the time passes the trends change the fashion change the people taste change. Hence with proper research to numerous kinds of recommendation systems. Users' listening habits have been impacted by the emergence of streaming apps and the use of music recommendation algorithms. You can create and develop algorithms that concentrate on the period during which a person listens to a certain music. This allows us to create custom playlists based on genre, mood, and timing requirements.

As music and sound have been useful in both physical and mental health as some sounds frequencies a promotes healing and making a person calm and relaxed with some particular sounf frequencies only.

Hence to conclude, a song or a music or a audio or just a sound is very powerful and can be used in any field from music industry for entertainment to medical purposes for healing. Researching in this topic has thought me allot about audio and music from application to visualization of sound frequencies this topic is vast, and I have insufficient knowledge, but I am willing to learn and research more and come up with ideas to contribute something meaningful in this field.

References

Bob L. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. (2013). Department of Architecture, Design and Media Technology Audio Analysis Laboratory

Schedl, M., Zamani, H., Chen, CW. et al. Current challenges and visions in music recommender systems research. Int J Multimed Info Retr 7, 95–116 (2018). https://doi.org/10.1007/s13735-018-0154-2