Music Genre Classification

Urmi Thakkar
200668299
Prof. Marcus Pearce
MSc Big Data Science

***Abstract*—Recently, Deep Neural Networks have proven to be very effective in classification tasks.A music genre,as described by wikipedia, is a category of music that identifies some pieces of music a shared tradition or a set of conventions.Music genre is distinguished based on musical style and form. Music Genre Classification is a challenging task especially now that there are so many large datasets available, making it even more challenging. In the last few years, a lot of studies have been conducted for the task of music genre classification using machine learning techniques. The traditional machine learning algorithms like SVM, logistic regression, K-nearest neighbor classification, etc were used to classify music genre and their performance was compared and evaluated against each other. However, recently, the deep learning models have become quite popular for classification tasks like image classification, audio classification and music genre classification. In this paper I present a comparative study between the performances of a deep learning architecture and various machine learning algorithms for music genre classification. The classification is performed using the very popular GTZAN Dataset.**

***Index Terms*—Music, Music Genre Classification, Machine Learning, Deep Learning, MFCCs, Melspectograms, GTZAN Dataset.**

I. Introduction

Music is the such an import part of our lives. Music genres are equally important as they can help elicit emotions and feelings. Genre can be described as a category of music based on several criterias and conventions. Genre of the music can play an important role in changing the mood of the listener, for example, genres like pop, hiphop can get you an energetic and so on. With the advent of various music applications like spotify, people can now search or play music according to the genres. This can be said to have such a positive impact on the listerner because the can select what type of songs they want to listen to at that point of time. It plays an important part in entertaining the listener with soothing and mood-appropriate music.

Music Genre Classification is a fundamental problem and is very challenging. Numerous models and algorithms have been developed over last few years to efficiently classify music according to its genres but some problems still exist. Various large datasets with Gigabytes of data are available for the task of music genre classification. However, because these datasets are huge, its a must to develop machine learning or deep learning models that are scalable and can classify music information based on several important criterias such as genre, lyrics, instrument, artists etc.

Music classification tasks have two main challenges- (a) Proper pre-processing of the audio data and extracting all the relevant audio features required to classify the music. (b) Selecting the most efficient machine learning or deep learning model in order to get the desired results in music classification. Rajanna et al. (2015). In this paper, we will see if machine learning models perform well in learning these important hand crafted features and classifying music according to their genres or the deep learning models outperform the traditional machine learning techniques.

Over the past few years, deep neural networks have shown to produce outstanding results in various fields such as image processing, speech recognition, audio processing as well as natural language processing. Deep learning models like Convolutional Neural Network, Recurrent Neural Network and others have proven to be efficient with enormously large datasets, producing the desired accuracies in the results. While machine learning techniques suffer because of the large size of the datasets and metadata, development of deep neural networks has proven to be very effective in dealing with the enormity of the datasets. Furthermore, in the domain of audio and music processing, deep neural networks have also shown the ability to perform improved audio feature extraction than the machine learning models, thus producing excellent results.

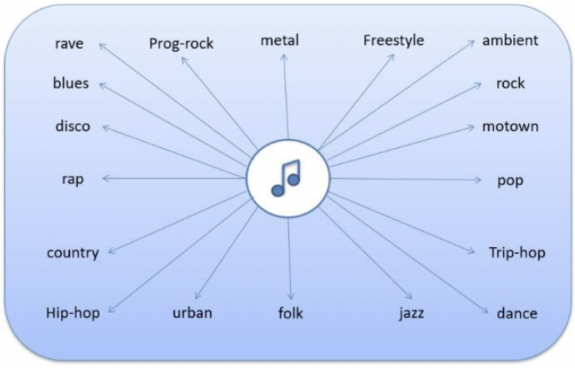

Fig. 1. Different types of music genres

Fig. 1. above shows the different types of genres that music can be categorised into. In this paper, for the task of music genre classification, we have compared two deep neural networks (a) Convolutional Neural Network and (b) VGG16 architecture. We have used these two architectures

to experiment on the GTZAN dataset and compared the performances of both models using metrics like accuracy and loss curves.

The paper is organised as follows:In section II we present the related work and advances in music genre recognition is discussed. In section III, the dataset structure is discussed.

II. Related Work

Over the years, many studies have been conducted Music Information Retrieval to describe and extract relevant features of an audio signal. Extracting relevant features is the most important task in music information retrieval and can boost the performance of music classification.Thus, the most existing research involves identifying and extracting relevant features and music representations in order to improve the performance of the model. Feng et al. (2017) have proposed a hybrid architecture consisting of parallel Convolutional Neural network focusing on spacial feature extraction and Bi-directional RNN blocks focusing on temporal frame orders feature extraction. The GTZAN dataset has been used for this task and final accuracy was 0.92. (Jakubík 2017) evaluate two Recurrect Neural Networks with gating, (i)Long=Short Term Memory (LSTM) and (ii)Gated Recurrent Unit (GRU). In the experiment, four datasets were used- Emotify, GTZAN, LAstFM and Ballroom. (Rafi et al. 2021) analyse and comapare improved versions of three deep learning models for music classification. The three models used are Bottom-Up Broadcast Neural Network (Liu et al. 2020), Independent Recurrent Neural Network (Wu et al. 2018) and a hybride Convolotional - Recurrent Neural Network in Time and Frequency dimensions(Wang et al. 2019).

(Goel et al. 2014) performed music genre classification using neural networks encorporating features like energy, loudness, beats, tempo, valence, danceability, speechiness, discrete wavelet transform. A musical representation of all these features was fed to a Multi-layer Perceptron network and the music was classified into two genres, Sufi and Classical, with an accuracy of 0.85. (Elbir et al. 2018) discuss some very important audio features used for music genre classification. The features like zero-crossing rate, spectral contrast, spectral centroid, MFCCs etc and max accuracy of 0.65 was produced. (Holzapfel & Stylianou 2008) proposed non negative matrix factorization based features for the task of music genre classification. A spectogram was factorized using the matrix factorization and fed to the classification model.

(Rajanna et al. 2015) compare the performance of machine learning models like SVM and l1 SVM along with some audio features with that of deep learning model. The metric used for comparison is the accuracies of the models. Thus, we can see that Music Genre Recognition is a challenging task and several research has been conducted in order to get the desired results.

III. Methodology

A. Dataset

In this paper we have used the GTZAN Dataset for the task of classifying music into its genres. The dataset is divided into 10 genres, with 100 audio files for each genre. All of the audio clips have a length of 30 seconds. The audio clips are 22050 with mom 16 bit audio file in .wav format. The GTZAN dataset is the most widely used dataset for evaluation in music genre recognition(MGR).The dataset is divided into genres like pop, classical, hip-hop, Blues, Country, Disco, Jazz, Metal, Reggae and Rock (also shown in Fig. 2. above).

1	Blues
2	Classical
3	Country
4	Disco
5	Hip hop
6	Jazz
7	Metal
8	Pop
9	Reggae
10	Rock

Fig. 2. Types of music genres in GTZAN music Dataset

B. Pre-processing

The first step was to read all the audio files, after reading the audio files, each audio file is now 30 seconds long. Next, we split each audio file into duration of 3 seconds each. Now, for the computer to relate each segment with one another, we have included 50 percent of the duration of the previous segment and 50 percent of the duration of the next segment. This will help the computer understand the relation between each segment.Now the audio segments are ready for the next step.

C. Feature Extraction

Feature extraction refers to the process of tranformation of raw audio into numerical features which can be fed to machine learning or deep learning models to perform the

desired task.Audio features are descriptions of sound. Different features capture different aspects of sound.

These features play a major role in training machine learning algorithms to recognise patterns in order to solve a particular task. In the process of feature extraction, the raw data is preserved in the original dataset.For our deep learning models we have extracted a feature called as Short term fourier Transform(STFT). Whereas, for our machine learning algorithms, we have used several handcrafted features from different domains listed below.

- Short Term Fourier Transform: It is an important feature as it enables us to extract spectograms which we can then feed our neural network. Fourier transform is not performed on the whole audio signal, but rather we apply it on small segments of the audio signals.So, STFT is calculated for all the segments of the audio signal, this can be done by applying windowing to the signal. The windowing fucntion takes the original signal and then multiplies it by a window function sample by sample. Each chunk of the audio signal after windowing is called as a frame. Windowing of samples is done using three parameters, (i) window size: window size is the amount of samples that we apply windowing to. (ii)Frame size: Frame size refers to the number of samples that we consider in each chuck of a signal.(iii) Hop size: Hop size tells us how many samples we slide to the right when we take a new chuck/frame. So STFT is applied to one chunk first and then it moves to the other until the end of the audio signal.

$$x_w(k) = x(k) \cdot w(k)$$

Fig. 3. The windowing function of STFT

The numeric formulation for STFT is given below.The result that we get from the Short term fourier transform is the fourier coefficient for the *k*th frequency at the *n*th temporal frame. The output of STFT is 2D array that has number of frequency bins and number of frames. It has reference to both frequency and time.

$$S(m,k) = \sum_{n=0}^{N-1} x(n+mH) \cdot w(n) \cdot e^{-i2\pi n\frac{k}{N}}$$

Fig. 4. The numeric formula of STFT

Now, the visualisation of STFT is done in form of spectrograms. Spectrograms have Time on the x-axis and Frequency on the y-axis. It shows how the different frequency bins evolve over time across different frames present in the original signal. These spectrograms are then fed to our deep learning models to perform classification.

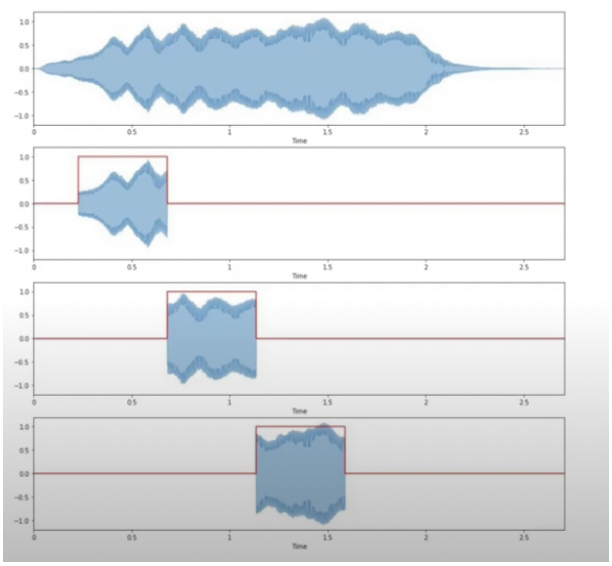

Fig. 5. Applying STFT to every frame of the audio signal.

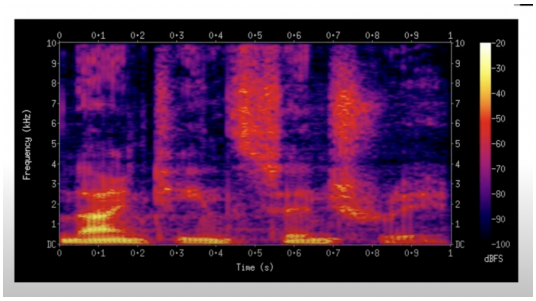

Fig. 6. Spectrogram

- **Time Domain Features**
  - Amplitude envelope:
    Amplitude envelope is the maximum value of all samples in a frame. It gives us a rough idea of the loudness. It is one of the features used for genre classification.
  - Zero-crossing rate:
    Zero Crossing rate is the number of times a signal crosses the horizontal axis.
  - Mel-spectrograms:
    Mel spectrograms are ideal for three reasons, time-frequency representation, human perceptual amplitude representation and human perceptual frequency representations. Humans perceive frequency/amplitude representations logarithmically, which is not possible to achieve with normal spectrogram, therefore mel-spectrograms come in handy.
  - Mel-frequency cepstral coefficients- MFCCs:
    Mel-frequency cepstral coefficients (MFCCs) are very good describers of music and capable of capturing the timbre. We usually take first 12-13 coef-

ficients in consideration because those are the most relevant ones. MFCCs are advantageous because they describe the large structures of the spectrum cutting down the noise that comes with the spectrum.

- Delta and delta-delta MFCCs:
  These are in-short the first and the second derivatives of the MFCCs. They are very important because they tell us how the MFCCs change over time in an audio file.
- Root-mean square energy:
  Root mean square energy is the square root of mean of sum of energy for all the samples of a frame.It is the indicator of loudness and therefore is a great feature for music genre classification. The energy signal is calculated as shown in Figure 7.

$$\sum_{n=1}^{N} |x(n)|^2$$

Fig. 7.

After calculating the energy, we can compute root mean square by using the formula in Figure 8

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

Fig. 8.

- **Frequency Domain Features**
  - Band-energy Ratio:
    Band energy ratio is (the sum of power in lower frequencies) / (sum of the power of higher frequencies). The split frequency gives us the threshold. All the frequencies above the threshold are higher frequencies and all the frequencies below the threshold are lower frequencies.It is extensively used in music genre classification.
  - Spectral Centroid:
    Spectral centroid is the weighted mean of the frequencies. It is one of the key frequency domain feature. It provides us with the center of gravity of the magnitude spectrum i.e. it gives us the frequency band which has most of the energy. It measures the brightness of the sound as to how bright or dull a certain sound is. It can be calculated as,

$$f_c = \frac{\sum_k S(k)f(k)}{\sum_k fk},$$

Fig. 9.

  - Bandwidth:
    Bandwidth is related to the spectral centroid. It is the range which is of interest i.e the spectral range around the centroid. It is the weighted mean of the distances of the frequency bands from the spectral centroid. It is extensively used in traditional ML based music genre classification.Bandwidth is calculated as,

$$[\sum_k (S(k)f(k) - f_c)^p]^{\frac{1}{p}}$$

Fig. 10.

  - Spectral Roll-off:
    Spectral Roll-off is the value of frequency below the threshold of the total energy in the spectrum lies.
- **Rythmic Features**
  - Tempo It is the estimation of music tempo to be used as a feature for audio classification. Tempo is one of the suitable features for audio genre classification because each genre will have a different playing speed.
- **Pitch content Features**
  - 1-Chroma features These features give us more information about the notes in the music played. They can therefore be very useful in genre classification because songs of different genres would have different patterns of notes.

D. Architectures

- **VGG16 Architecture**

  VGG16 (Simonyan & Zisserman 2014) is a convolutional neural network (CNN). A convolutional neural network is an artificial neural network which has an input layer, an output layer and a number of hidden layers. VGG16 is an object detection and classification algorithm which has an ability to classify 1000 images of 1000 different categories (Simonyan & Zisserman 2014). VGG16 has

also shown promising results in the domain of audio and music classification. The VGG16 architecture has 16 layers that have weights. It has thirteen convolutional layers, five max pooling layers and three dense layers. The input to the architecture is a 224 x 224 image with 3 RGB channels. VGG16 is one of the best models because its focuses on small number of hyper-parameters like having convolutional layers of 3 x 3 filter with stride = 1 with same padding and max pooling layer of 2 x 2 filter with stride = 2. The first Conv layer (Conv1) has 64 filters, the second has 128 filters, the third has 256 filters, the fourth and fifth conv layers have 512 filters. This stack of convolutional layers is followed by three fully connected layers and the final layer is the Softmax layer (Simonyan & Zisserman 2014). In this paper the VGG16 model used was proposed by (Ahmad* & Sahil 2019). The activation function used is RELU activation function. In the last layer, we have used the softmax function which assigns probabilities to each class in a multi-class classification problem and these probabilities add up to 1.0.

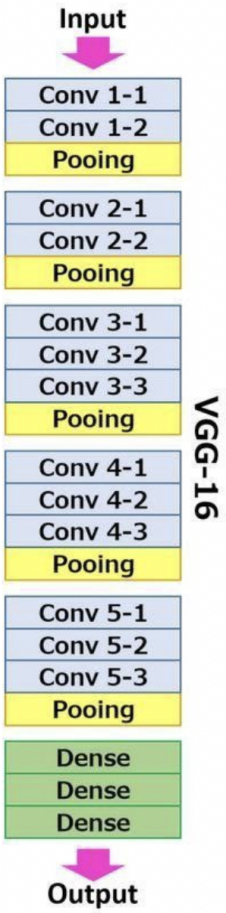

Fig. 11. Layers of VGG16 Architecture

Image retrieved from https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918#: :text=with%20transfer%20learning.-,VGG16%20Architecture,layers%20i.e.%2C%20learnable%20parameters%20layer. on 19/08/2022

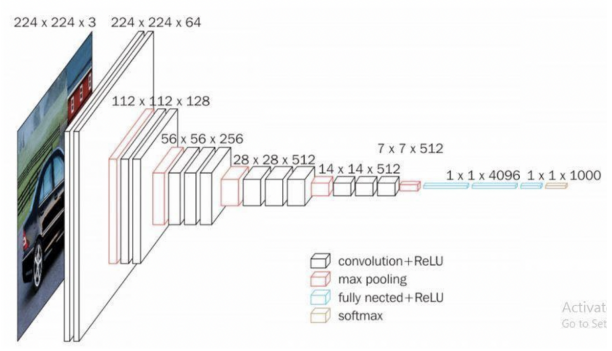

Fig. 12. The VGG16 Architecture

Image retrieved from https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918#: :text=with%20transfer%20learning.-,VGG16%20Architecture,layers%20i.e.%2C%20learnable%20parameters%20layer. on 19/08/2022

- **Support Vector Machine**

  Support Vector Machine (SVM) is a supervised machine learning algorithm which produces significant accuracy and requires less computation power. When given a set of labelled training data, SVM categorizes the test data into the target classes. The main aim of the support vector machine is to find a hyperplane in an N-dimensional space to classify the points in the data. SVM chooses the extreme points to create a hyperplane. Hyperplanes are the boundaries that help us to classify the data points. Its dimension depends on the number of classes the data needs to be classified into. For example, if there are 3 classes, the hyperplane is two-dimensional.
  The main advantage of a support vector machine over neural nets is higher speed and better performance but on a limited amount of data.

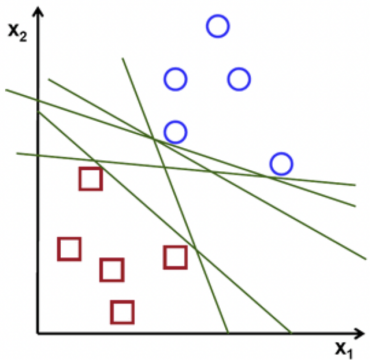

Fig. 13. Possible Hyperplane of SVM

Image source:
https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

- **Decision Tree Classifier**

  The decision tree classifier is a supervised machine learning algorithm which uses a set of rules to make some decisions. The logic behind decision trees is that the dataset features are used to create yes/no questions and the dataset is split until all the data points belonging

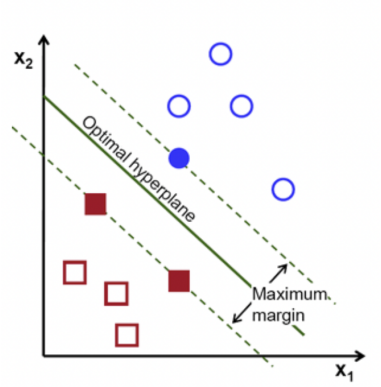

Fig. 14. Possible Hyperplanes of SVM

Image source:
https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

to each class are isolated. The data is organised in a tree like structure where internal nodes are the features of a dataset, the branches are the decision rules and each leaf node is the output. This structure has nodes which are the yes/no nodes of a question. The topmost node is called the Root node and the nodes in the bottom-most layer are called as leaf nodes. The goal of the decision tree is to create a training model that can be used to predict the class of a target by learning simple decision rules from prior data. Two main characteristics of a decision tree are (i) It mimics human decision making ability and therefore it can be easily understood. (ii) The tree like structure makes it easily understandable.

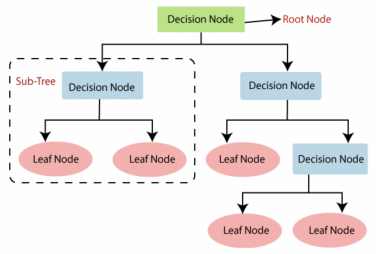

Fig. 15. Decision Tree Classifier

Image source: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithms

- **Random Forest Classifier**

Random Forest Algorithm is a machine learning algorithm that combines the algorithms of multiple decision trees to produce a single output. We learnt about decision trees in the above section, decision trees can be prone to bias and over-fitting. However, using multiple decision trees can mitigate these problems. They can produce more accurate results when the individual trees have no correlation with each other. It uses bagging and feature randomness to create a forest of uncorrelated decision trees. Feature randomness creates a random subset of features ensuring low correlation among the decision trees. Random Forest has reduced risk of over-fitting and is a very flexible model but it also has some challenges like its time consuming, is more complex and requires more resources.

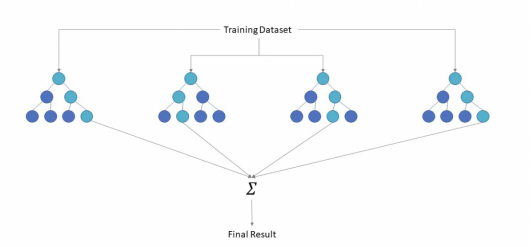

Fig. 16. Random Forest Classifier

Figure retrieved from https://www.ibm.com/cloud/learn/random-forest on 19/08/2022

- **Logistic Regression**

Logistic Regression is a supervised machine learning algorithm for classification. It is useful for data with numeric input variables and categorical target variable. Logistic Regression estimates the probability of occurrence of an event based on input variables.It is calculated as the probability of success divided by the probability of failure. Logistic regression has two types- (i) Binary Logistic Regression and (ii) Multinomial Logistic Regression. In binary logistic regression, the dependent variable has only two possible outcomes (0 or 1). In multinomial logistic regression, there are three or more classes that means the dependent varaible has three or more possible outcomes. Logistic Regression uses the Sigmoid function to squeeze the values in range and make an S curve graph as shown below.

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}.$$

Fig. 17. The Sigmoid Function

Image retrieved from https://medium.com/mlearning-ai/multinomial-logistic-regression-adb2a76eedcf

Fig. 18. Logistic Regression Curve

Image retrieved from https://medium.com/mlearning-ai/multinomial-logistic-regression-adb2a76eedcf

E. Evaluation

- **Metrics** Metrics are used to evaluate the performance of the models. There are various metrics that can be used to evaluate the performance of deep learning and machine learning models, we will be using:
  - Accuracy:
    Accuracy of a model refers to the percentage of correctly classifies samples from our test data.
  - Confusion Matrix:
    Confusion matrix is not necessarily a metric but its a tabular representation of the results of our model. It helps us understand the areas of strengths and weaknesses of our model on the test data.
  - F1 Score:
    F1 score is a weighted harmonic mean of Precision and Recall, the best score is 1.0 and the lowest score is 0.0. We will be using the weighted average of F1 to compare our machine learning classification models.
  - Classification Report:
    Classification report is used to understand the quality of the predicted outputs for a classification algorithm. It includes Precision and Recall. Precision is the ratio of true positives to the sum of true positives and false positives. Recall can be defined as the ratio of true positives to the sum of true positives and false negatives.

- **Results and Discussion**
  In this section, we will be discussing the modelling approaches and the results of the models discussed in Section III-D. The models will be evaluated based on the metrics discussed in Section III-E.
  First, we divide the data from the GTZAN dataset into training, validation and test data. The training data is then fed to our machine learning models along with the hand-crafted features mentioned in section III-C. To train the deep learning model, we've trained it on the spectrogram extracted from the training data. The spectrograms for the models were extracted using Short Term Fourier Transform as explained in Section III-C. This spectrogram data is then reshaped and fed to the VGG16 architecture. The architectures are first trained and then validated on the validation data and training and validation accuracies are evaluated. Once training and validation is done, we then test the models on the test data. The models are evaluated based on the accuracies and F1 scores produced on the test data. We then plot the confusion matrices to properly evaluated the areas of strength and weaknesses of our models.

Table I shows the final accuracy and F1- score of each model on the test data. Among the machine learning models which were trained using several hand-crafted features, Random Forest Classifier proved to out-perform all the other machine learning classifiers. Several experiments were conducted

TABLE I
COMPARISON OF PERFORMANCE OF THE MODELS ON TEST DATA

Models	**Metrics**	
	Accuracy	***F1-Score***
Decision Tree Classifier	0.45	0.44
Random Forest Classifier	0.60	0.59
Support Vector Machine Classifier	0.54	0.53
Logistic Regression	0.57	0.56
VGG16 Architecture	**0.83**	

i.e. hyper-parameter tuning was implemented by tuning the n_estimators to 100,200,300 but n_estimators= 300 performed the best. For Support vector machine, we experimented with the value of C and it was observed that as we decreased the value of C, the accuracy decreased as well. Below are the Confusion matrices of the machine learning models

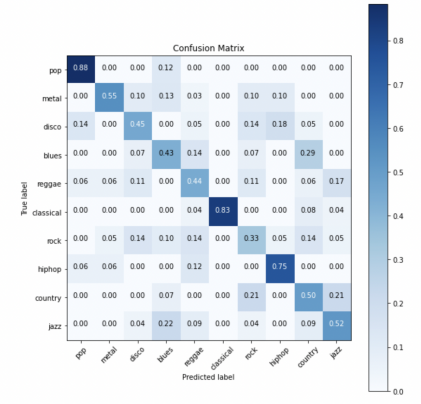

Fig. 19. Logistic Regression

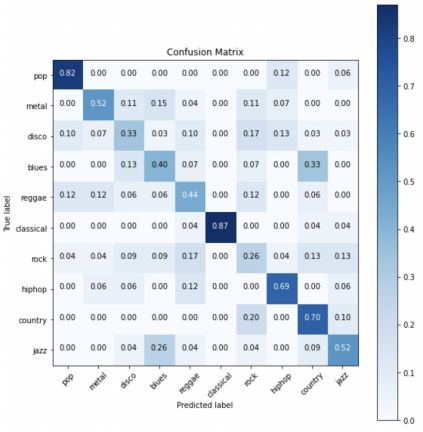

Fig. 20. Support Vector Machine

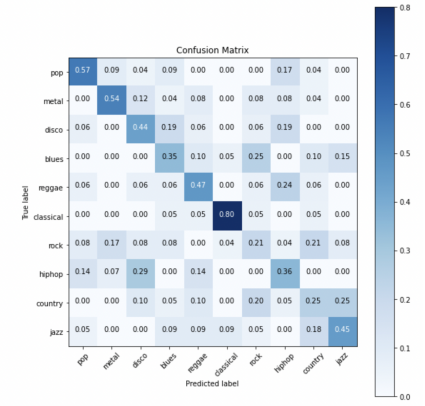

Fig. 21. Decision Tree Classifier

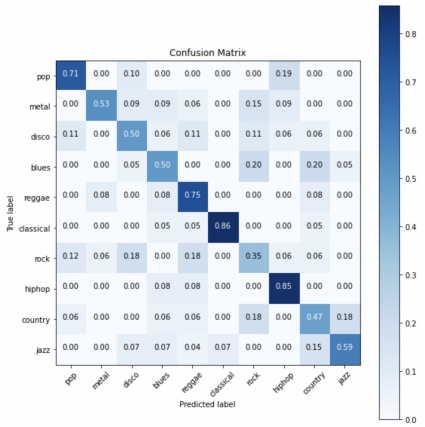

Fig. 22. Random Forest Classifier

As we can see, VGG16 architecture performs the best with the accuracy of 83%. It uses only the spectrogram to predict the genre of the music. This shows that a convolutional neural network architecture can improve the scores on a classification task like this.

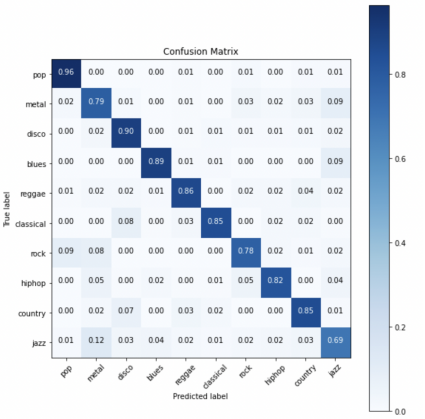

Fig. 23. VGG16 Architecture

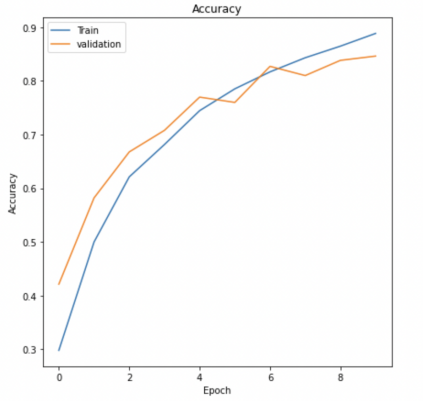

Fig. 24. Accuracy Curve for VGG16

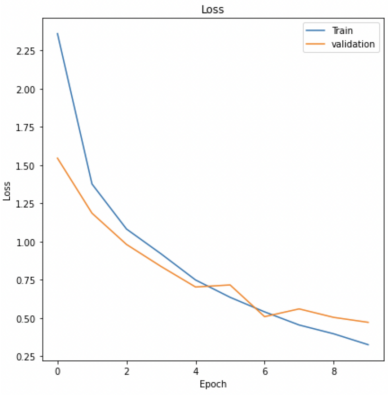

Fig. 25. Loss Curve for VGG16

F. Conclusion and Future Work

In this paper, we compared traditional machine learning algorithms with the popular deep neural network which is the VGG16 Architecture. We performed feature engineering by feeding some hand-crafted features to our machine learning models. The neural net was fed the spectrograms extracted using the Short-term-fourier-transform audio feature. Both the deep learning and machine learning are evaluated based on accuracy and confusion matrix representation is used to visualise the strengths and weaknesses of each model. To conclude, VGG16 architecture outperforms all the machine learning architectures with the final accuracy of 83% for 10 epochs on the test data. For future work, other neural networks can be implemented to produce improved classification of music genres. 2. VGG16 architecture has 138 million parameters which causes exploding gradient problem. Different ways can be studied to overcome VGG16's exploding gradient problem and further study can be conducted. Due to lack of resources and time, we could not perform proper hyper parameter tuning, therefore by hyperparameter tuning, accuracy can be increased in future.

REFERENCES

Ahmad*, F. & Sahil (2019), 'Music genre classification using spectral analysis techniques with hybrid convolution-recurrent neural network'.
URL: *http://dx.doi.org/10.35940/ijitee.A3956.119119*

Elbir, A., Cam, H., Iyican, M., Ozturk, B. & Aydin, N. (2018), Music genre classification and recommendation by using machine learning techniques, pp. 1–5.

Feng, L., Liu, S. & Yao, J. (2017), 'Music genre classification with paralleling recurrent convolutional neural network'.

Goel, A., Sheezan, M., Masood, S. & Saleem, A. (2014), 'Genre classification of songs using neural network', *2014 International Conference on Computer and Communication Technology (ICCCT)* pp. 285–289.

Holzapfel, A. & Stylianou, Y. (2008), 'Musical genre classification using nonnegative matrix factorization-based features', *Trans. Audio, Speech and Lang. Proc.* **16**(2), 424–434.
URL: *https://doi.org/10.1109/TASL.2007.909434*

Jakubík, J. (2017), Evaluation of gated recurrent neural networks in music classification tasks, *in* 'ISAT'.

Liu, C., Feng, L., Liu, G., Wang, H. & Liu, S. (2020), 'Bottom-up broadcast neural network for music genre classification', *Multimedia Tools and Applications* **80**(5), 7313–7331.
URL: *http://dx.doi.org/10.1007/s11042-020-09643-6*

Rafi, Q. G., Noman, M., Prodhan, S. Z., Alam, S. S. & Nandi, D. (2021), 'Comparative analysis of three improved deep learning architectures for music genre classification', *International Journal of Information Technology and Computer Science* **13**, 1–14.

Rajanna, A. R., Aryafar, K., Shokoufandeh, A. & Ptucha, R. (2015), 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.

Simonyan, K. & Zisserman, A. (2014), 'Very deep convolutional networks for large-scale image recognition'.

Wang, Z., Muknahallipatna, S. S., Fan, M., Okray, A. & Lan, C. (2019), 'Music classification using an improved crnn with multi-directional spatial dependencies in both time and frequency dimensions', *2019 International Joint Conference on Neural Networks (IJCNN)* pp. 1–8.

Wu, W., Han, F., Song, G. & Wang, Z. (2018), Music genre classification using independent recurrent neural network, *in* '2018 Chinese Automation Congress (CAC)', pp. 192–195.

MSc Project - Reflective Essay

Project Title:	Music genre Classification
Student Name:	**Urmi Thakkar**
Student Number:	**200668299**
Supervisor Name:	Prof Marcus Pearce
Programme of Study:	MSc Big Data Science

Music genres can be described as categories of music classified using a set of characteristics such as pitch, tempo, rhythm, power and harmonic content. (Poria, et al., 06/29/13) Music genre classification can be used to retrieve music from large collection of music on the web. It is a challenging task and has attracted a lot of attention in the area of Music Information Retrieval (MIR). In this paper, we have compared machine learning and deep learning approaches for music genre classification. We performed feature engineering to extract the relevant features from raw audio data and used these features to train our models. We have compared four machine learning models, (i) Decision tree Classifier, (ii) Support Vector Machine (iii) Random Forest Classifier (iv) Logictic Regression and one deep neural network- VGG16 Architecture. We compared the performances of these models on the test data based on accuracy and F1-scores. Let's discuss some strengths and weaknesses of the models,

1 Analysis of Strengths and weaknesses:

1.1 Strengths:

- In the process of feature extraction, we have used several handcrafted features like amplitude envelope, pitch, spectral centroid, spectral roll-off, bandwidth, tempo and a few more. Due to all these features, the machine learning models can learn and classify the audio efficiently producing desired results.
- We have extracted spectrogram from the raw audio data and these spectrograms are then fed to our neural network. Spectrograms have Time on the x-axis and Frequency on the y-axis. It shows how the different frequency bins evolve over time across different frames present in the original signal.
- By finding patterns in spectrogram, machine learning models are able to extract dominant audio per time frame in a waveform.
- Convolutional neural networks boost the accuracy of classification because it can extract relevant audio features.

Weaknesses:

- VGG16 architecture has 16 layers and therefore it is very slow to train.
- 138 million parameters cause exploding gradient problem.
- Machine learning algorithms are fast and require less computation power but their performance degrades with increase in the size of a dataset. Therefore, machine learning models are preferred when the data in limited in size.
- As classifying genre into exact different classes is a challenging task, performance of SVM degrades because SVM works well when there classes that can be completely distinguishable.
- Due to lack of resources, we could only implement 10 epochs as VGG16 is very slow to train.

Presentation of possibilities for further work:

Further studies can be conducted to improve the task of Music Genre Classification, such as:

1. Other neural networks can be implemented to produce improved classification of music genres.
2. Further study can be conducted and solutions can be found to overcome VGG16's exploding gradient problem,
3. Future studies may include research on ways to eliminate noise in the audio data so as to improve the efficiency of the models in classifying the audio data and achieve better performance.
4. Due to lack of resources and time, we could not perform proper hyper parameter tuning, therefore by hyperparameter tuning, accuracy can be increased in future.
5. In the future, I intend to study other network architectures such as Long Short term Memory (LSTM) or Generative adversarial network (GAN) for music genre classification.
6. I intend to explore other audio features that would be relevant in classifying music genres.
7. Also, other representations can be studied to better understand the details and performances of machine learning and deep learning models,

3. Critical analysis of the relationship between theory and practical work produced

While doing this project, I realised the importance of understanding the theoretical concepts behind every function or algorithm. In the beginning, I faced a lot of difficulties in starting the project because I did not understand where to start from as I was new to this area of audio and music. I mostly concentrated on understanding practical implementations of models in the domain of music but I couldn't understand most of the things because my theoretical concepts were not clear. Then, I studied about the mathematical theory behind each feature of music and how it affects the audio signal. I learnt more about audio signal processing and that gave me so much clarity in this domain. Therefore, according to me there is a very strong relationship between theory and practical work produced because its very necessary to understand what goes behind and how the algorithm actually works in the backend.

References:

(Poria, et al., 06/29/13)In: Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Rodríguez, J.S., di Baja, G.S. (eds) Pattern Recognition. MCPR 2013. Lecture Notes in Computer Science, vol 7914. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978 (Poria, et al., 06/29/13) (Poria, et al., 06/29/13) (Poria, et al., 06/29/13) (Poria, et al., 06/29/13)-3-642-38989-4_26 (Poria, et al., 06/29/13)