20185116

by Divas Gupta

Submission date: 09-May-2022 02:41PM (UTC+0530)
Submission ID: 1831892875
File name: Finalsem_ProjectReport_Group4_1.pdf (1.36M)
Word count: 5496
Character count: 27940

A Project Report

on

MUSIC STYLE CLASSIFICATION USING CNN

Submitted in the Partial Fulfillment of the Requirements
for the award of

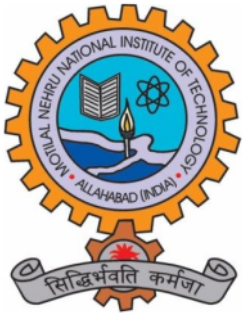
Bachelor of Technology
in
Electronics & Communication Engineering

By

Divas Gupta(20185116)
Priyanka Soni(20185053)
Dipesh Sharma Poudel(20185105)

Under the guidance of

Dr. Vinay Kumar Srivastava
Professor



Department of Electronics & Communication Engineering
Motilal Nehru National Institute of Technology Allahabad
Allahabad – INDIA

Department of Electronics & Communication Engineering
Motilal Nehru National Institute of Technology Allahabad
Allahabad – INDIA

CERTIFICATE

This is to certify that the work contained in the thesis titled "**Music genre classification using CNN**", submitted by **Divas Gupta**, **Priyanka Soni** and **Dipesh Sharma Poudel** in the partial fulfillment of the requirement for the award of Bachelor of Technology in Electronics and Communication Engineering to the Electronics and Communication Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, is a bonafide work of the students carried out under my supervision.

Date: 02 May 2022

Place: Prayagraj

Dr. Vinay Kumar Srivastava
Professor
ECE Department
MNNIT, Allahabad

<Similarity Index Certificate from IPR Cell>

Acknowledgement

We take this opportunity to express our deep sense of gratitude and hearth felt thanks to our project supervisor, **Dr. Vinay Kumar Srivastava**, Department of Electronics & Communication Engineering, Motilal Nehru National Institute of Technology, Allahabad for his constant guidance and insightful comments during the course of the work. We shall always cherish our association with him for his constant encouragement and freedom to thought and action rendered to us throughout the work.

We are also thankful to our colleagues and friends for their constant support. Finally, we deem it a great pleasure to thank one and all that helped us directly or indirectly in carrying out this work.

Date: 02 May 2022

Place:Prayagraj

Divas Gupta(20185116)

Priyanka Soni(20185053)

Dipesh Sharma Poudel(20185105)

Abstract

Our study perform some comparison of different machine learning algorithms to automatically classify music into their genres. First, a review of existing strategies and methods is carried out, both in terms of feature engineering and the provided algorithms. Creating two different data sets from an online online music archive, using sampling and over-sampling to classify classes. Fifty-four features are extracted manually for each sample, using the Librosa and Sound Sound Analysis Libraries. Then two sets of analyzes were performed; initial comparisons, individual features, average values of each type; the second compares each element with the value of f (points calculated using the Scikit-learn's Select K Best function). Four stages were developed - a neural network, a support-vector machine, a random forest and a gradient magnification machine - and training and testing were performed individually, creating each set of two different data sets. The results show that the vector support system is the most suitable algorithm for the job, with Hip-Hop music being the most precisely divided genre. Finally, an evaluation of the entire process is performed, which includes an examination of the database selection, feature selection, and various aspects of project management.

Table of Contents

Certificate i

Similarity index Certificate from IPR Cell ii

Acknowledgement iii

Abstract iv

List of Figures vii

Abbreviations viii

Chapter 1: Introduction

1.1 Introduction 1

1.2 Motivation 2

Chapter 2: Literature Survey

2.1 Introduction to ML 3

2.2 Method in ML 3

2.3 Algorithms in ML 6

2.4 Applications of ML 8

2.5 Limitations of ML 10

Chapter 3: Methodology and Architecture

3.1 Dataset 11

3.2 Feature Extraction 12

3.3 Convolution Neural Network 14

3.4 Artificial Neural Network 15

3.5 K Nearest Neighbor 16

Chapter 4: Result and Performance Evaluation

4.1 Evaluation Metrics 18

4.2 Result 22

Chapter 5: Conclusion

5.1 Conclusion 25

5.2 Future Scope of work 25

References 26

List of Figures

Fig 2.1 Flow chart of supervised learning

Fig 2.2 Unsupervised learning

Fig 2.3 Pipeline of Reinforcement learning

Fig 2.4 Example of decision Tree

Fig 3.1 MFCC Plot

Fig 3.2 CNN Architecture

Fig 3.3 ANN Architecture

Fig 3.4 Flow chart of the process

Fig 4.1 Confusion Matrix

Fig 4.2 Example

Fig 4.3 Summary of KNN algorithm

Fig 4.4 Summary of random forest algorithm

Fig 4.5 Summary CNN algorithm

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolution Neural Network
DCT	Discrete Cosine Transform
KNN	K Nearest Neighbor
DBM	Deep Boltzmann Machine
EM	Expectation Maximization
GRU	Gated Recurrent Unit
MFCC	Mel Frequency Cepstral Coefficient
MSD	Million Song Dataset
LSTM	Long Short Term Memory
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SVM	Support Vector Machine

Chapter: 1

Introduction

1.1 Introduction

As per data 24,000 songs get uploaded every day on different platform some major platforms are Spotify, Apple, Google Music, Napster, Dezzer.
so its impossible for us to remember the genre of music with name of music so its very important to classify the music for user easy.
in our project we have taken 10 major genre for classification they are :- blues, classical, country, disco, hiphop, jazz, reggae, rock, metal, and pop.
Each genre contain 100 music and each music is 30 sec long.

Our dataset has two folder namely:-

1) Genres original :- A collection of 10 genres with 100 audio files each, all having a length of 30 seconds.
2) Images original :- A visual representation for each audio file.Since, it represented in form of images neural networks are best suited for the classification purpose . We can't give audio to neural network but what we can do is to convert those audio into their corresponding spectrograms and feed them into CNN. For that purpose this folder is useful.

Why genre classification:-

1) Nowadays, it is most common phenomenon that people have some tune in their mind but they can't remember the song name. So, music style classification comes into picture here by providing an aid to information retrieval systems which are based on music.
2) Our project automatic music style classification can proved to of great use in searching and classifying large electronic music libraries.
3) In order to generate better recommendation on the basis of different ratings and reviews given, genre or books metadata is used.
4) Sometimes, user wants to listen music for mental relief. This project can be of great use for this purpose. Our project can be integrated to a mood classification model. So, instead of identifying the music genre it can suggest music on basis of user's mood.

1.2 Motivation

As per data on a single day 60,000 songs streamed in a day. This is really huge almost one songs per second , many people use to be sad and they want to feel happy but they don't know the name of song so they can listen song from classification of happy.

Our project aim to give a systematic approach to store the music data as music data is too vast, we aim to help human sociological and psychological nature, our project aim to save time while selecting songs our project will help to quickly select songs irrespective of occasion, we want to diversify the music system and to help the publisher to define the type of music.

To detect the type of music our project can be used as it will give unbiased result as it don't have human intervention. Our main aim motivation is to make music a best friend of music users.

.

Chapter: 2
Literature Survey

2.1 Introduction to ML

Machine Learning (ML) has been developed from Artificial Intelligence, a field of computer science. Machine Learning (ML) is a multi-disciplinary, mathematical compilation and computer science algorithms widely used in predictable analysis and classification. In recent decades, the proliferation of Artificial intelligence (AI) has become a broad and exciting field in computer science as technology prepares machines for human performance, and aims to train computers to solve real-world problems with a high level of success. As we see the growth of science and technological advances AI systems are now able to learn and improve by using prior knowledge without explicit help code when exposed to new data. Ultimately it leads to machine learning technology (ML) that uses learning algorithms to read from available data. Since, the term mining is known to us which means extracting natural substances from earth's crust by digging deep into the earth. Also, Machine Learning uses the same concept of mining but here mining is related to data in order to extract information from large-sized databases. Machine learning and data mining algorithms are invested in various fields such as Computer networking, tourism and tourism industry, financial forecasting, telecommunications industry and power forecasting and more.

2.2 Methods used in Machine Learning

2.2.1 Supervised Learning

Contains a given set of input variables (training data) pre-labeled and targeted data. It uses a variable input to generate a map function to map the required output to the corresponding input. The parameter adjustment process continues until the system obtains the appropriate level of accuracy regarding the teaching data.
In other words Supervised learning is a machine learning method that maps out output to the desired inputs based on a pair of output inputs. Considering the work from the training data labeled which includes a set of training examples.

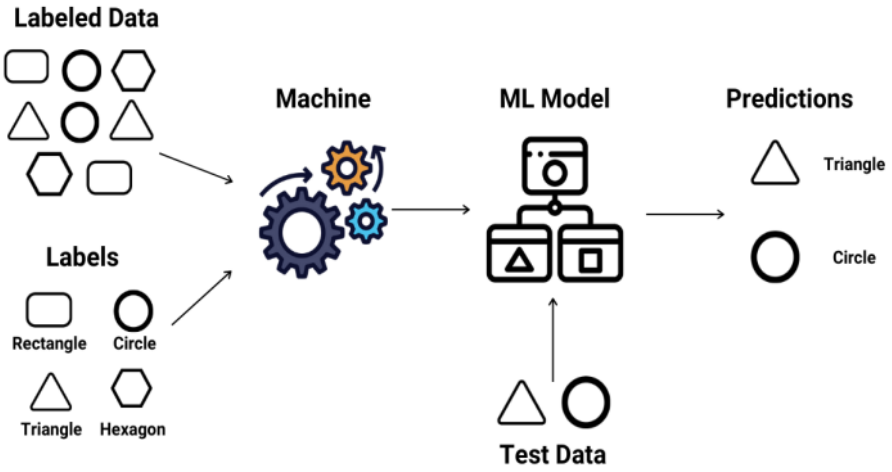


Fig 2.1: Flow chart of supervised learning

The supervised learning algorithm analyzes data given in the form of training and generates activity which we desired in the form of a function, which can be used to map new examples. The right mapping will allow the algorithm to accurately determine class labels in very new situations which are never seen before. This requires a learning algorithm to perform normally from training data to situations that do not appear in a "rational" way.

2.2.2 Unsupervised Learning

In this algorithm we only have training data rather than outcome data. That input data does not have a previous label. It is used for classifiers by identifying existing patterns or collections in input databases. In other words Unsupervised learning is a form of machine learning that looks at patterns that have never been seen in a data set that does not have pre-existing labels and has little human supervision. In contrast to supervised reading that often uses human-written data, Unsupervised Learning, also known as self organization, allows for modeling opportunities for overcrowding. It forms one of the three main stages of machine learning, as well as supervised and reinforced reading.

The two main methods used in Unsupervised learning are the principal component analysis and cluster(group) analysis.

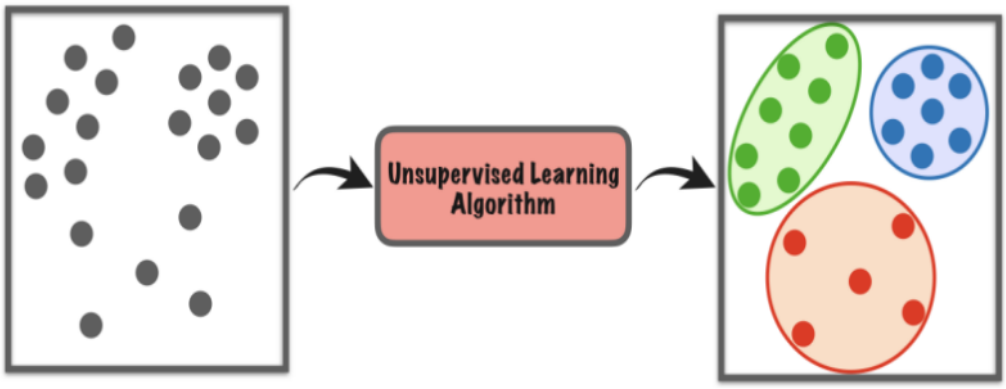


Fig 2.2 : Unsupervised learning

2.2.3 Reinforcement Learning

Using this algorithm machine is trained to map the action to a specific decision which is why the prize or response Symbols are generated. The machine is trained to detect the most rewarding actions by rewards and punishment using previous sensations.

It is a machine learning area about how software agents should take steps in place to develop a vision for the accumulated reward. Reinforcement learning is one of the three basic methods of machine learning, next to supervised reading and non-supervised learning.

Reinforced Learning differs from supervised learning that do not requires paired output/inputs, and require less appropriate steps that need to be clearly adjusted.

Nature is often referred to in the Markov (MDP) decision-making process, because many reinforcement learning algorithms in this context use flexible editing.

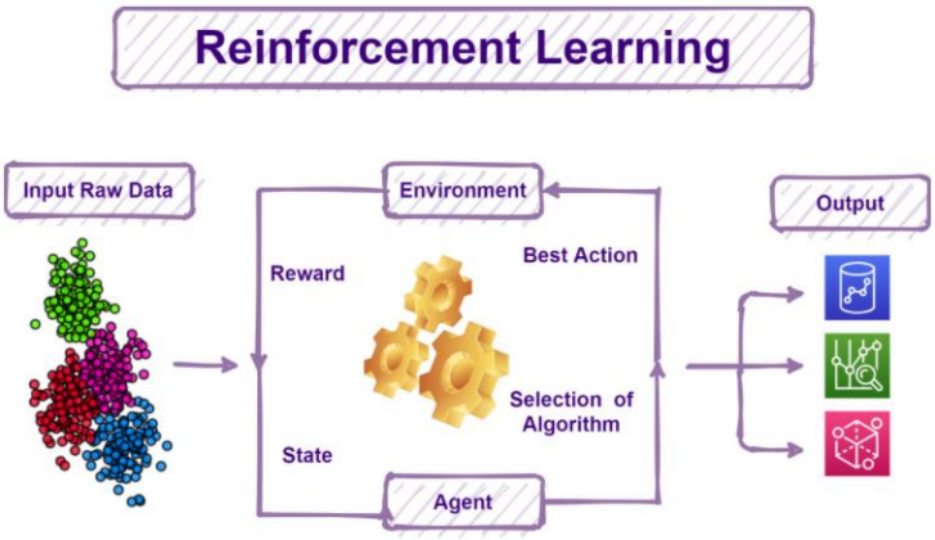


Fig 2.3 Pipeline of Reinforcement Learning

2.3 Algorithm of Machine Learning

2.3.1 Regression Algorithm

In Regression algorithms predictions are modeled by establishing the relationship between variables using error rate. Continuous value is predicted by the Regression strategy. Variables can be value, temperature. Regression algorithms are a type of Supervised algorithms. If we talk about features of this algorithm is that is pays attention on the relationship between the target output and the input features to predict the amount of new data. Algorithms based of regression produces output values on the basis of input feature of the data which is given to the system. Algorithm builds a model on training data features and uses that trained model to predict the amount of new data.

Popular areas of regression algorithms are as follows:

- Linear, Quadratic Regression algorithm
- Normal Decline of Small Squares
- Multivariate Adaptive Regression Splines
- Logistic Regression
- Moderately smooth scatter structure
- Step-by-Step Regression

2.3.2 Decision Trees

Decision trees are most of the used in classification problems like logistic regression. They split attributes in two or more groups by sorting them on the basis of their values. Each tree consists of nodes and branches. Attributes of the clusters are represented by

each node and branch represents its value. Pre-pruning and post-pruning are some techniques to improve their accuracy.

- Iterative Dichotomized 3
- Chi squared Automatic Interaction Detection
- C5.0and C4.5 (different versions of a powerful approach)
- Decision Stump
- Classification and Regression Tree
- Conditional Decision Trees

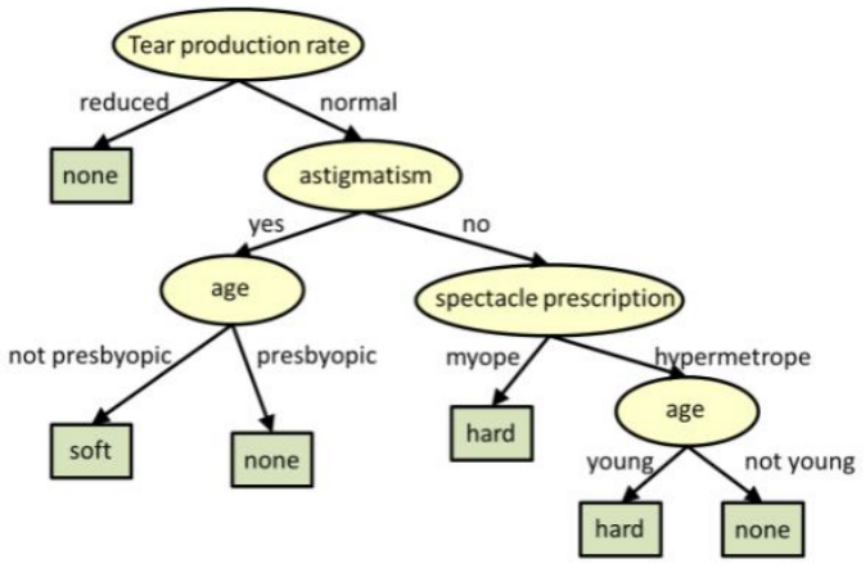


Fig 2.4 Example of Decision Tree

2.3.3 Bayesian Algorithms

Machine Learning is a variety of Computer Science fields like math and algorithm. The statistics control and measure uncertainty and are represented by Bayesian algorithms based on the theory of probability and Bayes theory.

The most common Bayesian algorithms are:

- Bayesian Belief Network (BBN)
- Gaussian Naive Bayes
- Naive Bayes

2.3.4 Clustering Algorithms

This algorithm divides objects into different types of collections. Divides objects into groups where each subset sets share a certain similarity. It is an unsupervised learning method and its methods are categorized as a domain or network clustering and partition clustering. The most popular clustering algorithms are that are used are listed below:

- K Means
- Expectation Maximization (EM)
- K Medians
- Hierarchical-Clustering

2.3.5 Deep Learning Algorithms

Deep Learning techniques called deep because they are made of many number of layers and each layer have much more neuron and are the further advancements made in Artificial Neural Networks . They are more complex neural networks containing lot of hidden layers are large in size. The widely used algorithms for deep learning are:

- Deep Belief Networks
- Deep Boltzmann Machine (DBM)
- Convolution Neural Network (CNN)
- Recurrent Neural Network (RNN)

Chapter 3:

Methodology and Architecture

3.1 Data Sources

The GTZAN Database, detailed in Tzanetakis (2002), is a well-known source in the field of MIR it contains 10 groups of 100 30-second song quotes, each group representing a single genre, a total of 1000 audio files in total. The database will be sorted into folders - one for each type - making it straightforward roaming, and the nature of the database balance makes it attractive machine learning projects.

However, despite these attractive features, GTZAN has many drawbacks. First, with only about 100 items per class, it is a relatively small database, in context in other machine learning projects. Deep learning is often regarded as thriving on a large amount of data (Ng, 2015), and having a few training examples means that class dividers have little knowledge on which to build their models. Therefore it makes sense to see if a large database will be available. In addition, Sturm (2013) identified a number of databases, including repetition of songs (i.e. different quotes from the same song that seem different tracks in the database) and the non-labeling of tracks (i.e. tracks labeled as wrong type).

For our project, one data source can't be relied on so, the second database to be considered was the Million Song Dataset (Bertin-Mahieux et al., 2011). As the name implies, MSD contains a million songs, which are distributed over a variety of species and subspecies. A million data points is a big data set. How and there are many good machine learning projects made with very few uses
examples of training.

However, MSD differs from other data sets such as GTZAN in importance method: while GTZAN is able to provide access to the audio files of the database itself (due to the ambiguity of the songs), MSD is collaborating with current celebrities music, meaning that access to audio files cannot be granted, for reasons of copyright. A million audio file databases will also take a few terabytes of data, making it difficult to download and store for most researchers. Instead, the database contains pre-released features and metadata, such dances as well thunder, of each song. Although thes may be interesting to analyze elsewhere context, it seems necessary to be able to access audio files themselves, to allow the release and analysis of similar audio features the timbre and pitch-content features described above. So, another database became required.

3.2 Feature Extraction

As mentioned earlier also , now our database consists a total of 1000 music files having 100 audios belonging to each genre in .wav and a combination of 10 styles.For each piece, we made samples a a 2-second window in four random areas, thus enlarging our data into 8000 clips for two seconds each.

We use 22050Hz as our sampling rate , this leaves us with 44100 raw audio input features.We took 20ms as our frame size which is used as window that translates the whole audio file. The length of the audio sample and the size of the feature space were found to be a perfect match at 44100 features. So, after pre-processing, we have (8000,44100), where each element represents an amplitude peak at a given point in 44100. For each of our cross validation and test sets, we used 100 samples of un-augmented data. Also, by converting audio files to mel-spectograms we can perform pre-processing.We've seen considerable improvements in performance across all models as a result of this. Mel-spectograms are a popular tool for analysing noise because they closely resemble how people perceive noise (i.e., log frequency). To turn raw sound into a mel-spectogram, Fourier transient transitions should be applied to smooth sound windows, which are usually around 20ms wide.With signal x [n], window w [n], frequency axis ω, and shift m, these are calculated as:

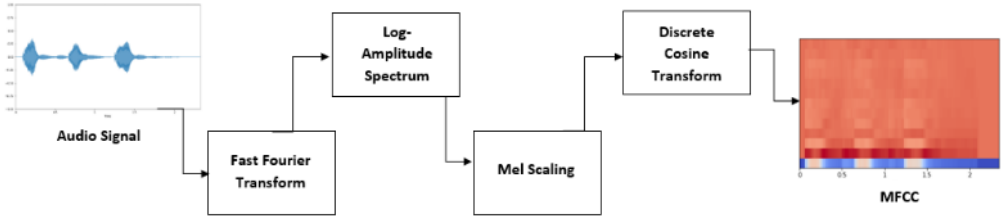
$$\text{STFT: } S(m,k) = \sum_{n=0}^{N-1} x(n+mH).w(n).e^{-j2\pi n\frac{k}{N}}$$

m: frame number

These are then transformed to the mel scale by mapping the frequencies f by:

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

Then we take the DCT of the result (common in signal processing) to obtain our MFCC(Mel frequency cepstral coefficient).



The resulting data can be visualized below:

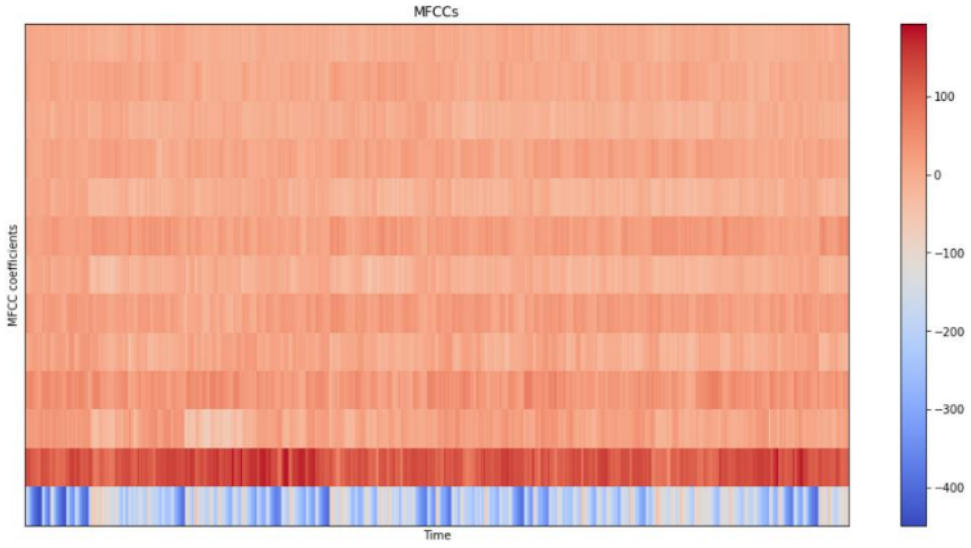


Fig 3.1 MFCC plot

3.3 Convolution Neural Network

In project, a convolutional neural network (CNN) is used to classify the music genre, contains convolution, pooling and fully-connected layers. Also, ReLU(Rectified Liner unit) f(x)=max(0,x) is a type of activation function which is used to add non-linearity. The ReLU works much quicker than conventional f(x) = tanh(x) or f(x) = $\frac{1}{1+e^{-x}}$ .Dropout is the technique use to deal with overfitting which randomly selects the neurons to 0 with probability 0.5. The neurons which are converted to zero to the forward pass and in back propagation.

The main function of CNN is to extract most relevant features from audio. VGG-16 and ResNet are common recommended as image encoders but we have built our own model for this.A brief summary of our CNN model is explained below with a well defined architecture of our model:

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_3 (Conv2D)            (None, 128, 11, 32)       320
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 64, 6, 32)         0
_________________________________________________________________
batch_normalization_3 (Batch (None, 64, 6, 32)         128
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 62, 4, 32)         9248
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 31, 2, 32)         0
_________________________________________________________________
batch_normalization_4 (Batch (None, 31, 2, 32)         128
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 30, 1, 32)         4128
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 15, 1, 32)         0
_________________________________________________________________
batch_normalization_5 (Batch (None, 15, 1, 32)         128
_________________________________________________________________
flatten_7 (Flatten)          (None, 480)               0
_________________________________________________________________
dense_26 (Dense)             (None, 64)                30784
_________________________________________________________________
dropout_10 (Dropout)         (None, 64)                0
_________________________________________________________________
dense_27 (Dense)             (None, 10)                650
=================================================================
Total params: 45,514
Trainable params: 45,322
Non-trainable params: 192
```

Fig 3.2 CNN architecture

3.4 Artificial Neural Network

Since, it is a artificial neural network, we have used fully connected neural network and, with our activation function as ReLU and 6 layers, and cross-entropy losses. Input in our neural network is 1-dimensional feature vector so to do this we flatten our 2-D Mfcc vector . For each layer, use it to activate ReLU work at the output of each node, according to the formula:

$$\text{ReLU}(x) = \begin{cases} x & ; x \geq 0 \\ 0 & ; x < 0. \end{cases}$$

In the output layer some kind of function is needed which can generate a 1-D vector corresponding to predictions of each class. In our case it will be a 10 x 1 vector. So, it was found that softmax is best suited for this purpose. It provides the probability associated with each class and then that class in this case it was genre which has maximum probability associated with it is taken as our predicted result:

$$\sigma_{j(z)} = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

To optimize our model, we minimized our loss function in our case it was cross entropy loss. The basic concept behind error minization is to calculate the first order derivative and taking those values which corresponds to minima. This is usually done by gradient descent method. Small steps are taken in the direction where gradient is descending to finally reaching to a minima. The loss function used in our case is as follows :

$$CE(\theta) = -\sum_{x \in classes} y(x) \log(\hat{y}(x))$$

Artificial Neural Network is prone to overfitting which means it gives high accuracy in training set but very low accuracy during testing time. To deal with it regularization techniques like L2 regularization or dropout are used. In case of neural network most widely used is dropout. L2 is best suited for regression problems as it adds a penalty term in loss function so if loss goes too low means accuracy goes too high it tries to control it by adding extra term in loss function thereby increasing it in case it goes very low.

Layer (type)	Output Shape	Param #
flatten (Flatten)	(None, 1690)	0
dense (Dense)	(None, 512)	865792
dense_1 (Dense)	(None, 256)	131328
dense_2 (Dense)	(None, 64)	16448
dense_3 (Dense)	(None, 10)	650

Total params: 1,014,218
Trainable params: 1,014,218
Non-trainable params: 0

Fig 3.3 ANN Model

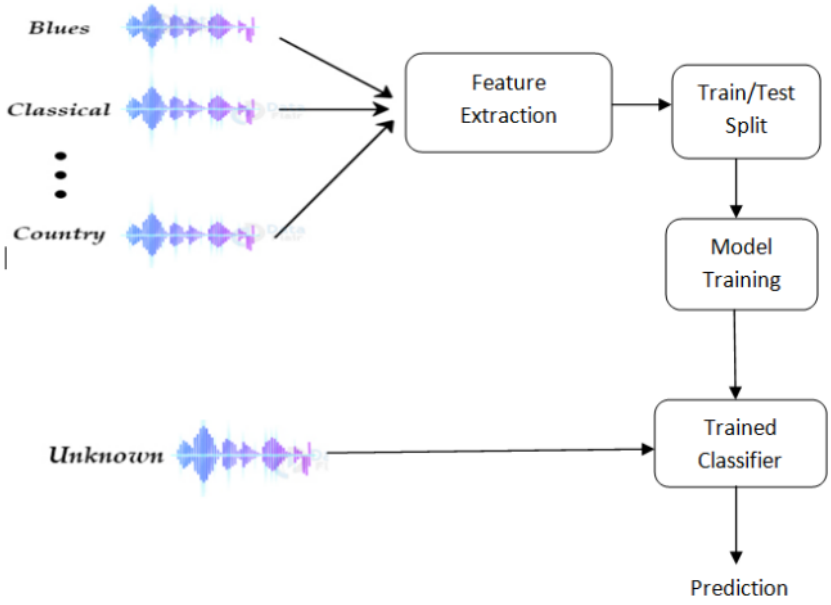


Fig 3.4 Flow chart of the process

3.5 K Nearest Neighbor

After reducing the size using PCA , we used an algorithm for neighbors close to k. In this algorithm we initialize our testing sample with its corresponding feature and measure the Euclidean of this particular sample with all the training and choose the samples which have least distance and among them took the most frequently occurring class.

Let some of our data points seems to look like this x (1), . . . , x(12) be the 112 closest neighbors to x, which will then return the greatest value on $\| x - x (i) \|_2$, where $\| \cdot \|_2$ denotes the Euclidean distance between points. Then w_i for each i is given below that's how we use it in our model

$$w_i \propto ||x - x(i)||, \qquad \sum_{i=1}^{12} w_i = 1.$$

Finally, we return

$$\arg max \sum_y w_i \ (y = y^{(i)})$$

the most frequent label in x's 12 nearest neighbors when calculated in regard to distance. This was implemented with scikit-learn.

Chapter: 4
Result and Performance Evaluation

4.1 Evaluation Metrics

The use of Evaluation Metrics is that after we are done with building a prediction based classification model, we need to evaluate the performance of the model, i.e. how well our trained model is performing in determing the outcome of new samples which are never seen before.

In other words we need to measure model accuracy with prediction errors using a new set of test data. The effect of observation are known to us with respect to our test data of a predictable model and can be evaluated by comparing predicted outcome values against known result values.Various types of evaluation techniques for classification model:-

- Average classification accuracy: It is representing the proportion of correctly classified observations using a model from a test dataset .
- Confusion matrix: It represents a 2x2 matrix which includes the amount of true positives, true negatives, flase negatives and false positives.
- Precision, Recall and Specificity: all of these 3 constitute the most commonly used performance indices of a prediction based classification model.
- ROC curve: A summary that is a picture of the overall performance of the model, which shows a portion of the real good and false positive of all the possible termination opportunities. The Area Under the Curve is used to know how well our model is working.

Introducing Confusion Matrix

The thing about Confusion matrix (also known as an error matrix) is that it can be seen as the summarized table/matrix which is of great help to analyse the results/prediction of a model based on classification. The number of positive and negative predictions is shortened by the calculation values.

We can say that the confusion matrix can be of dimension NXN which is used to verify our results , the N in size of matrix represents number of .It performs a comparison kingd of operation on the actual target values with those predicted by our respective model used for training. This gives us a complete idea of how well our classification model works and what types of errors it makes.

For a binary classification problem, so we would have a 2 x 2 array as shown below which have 4 values:

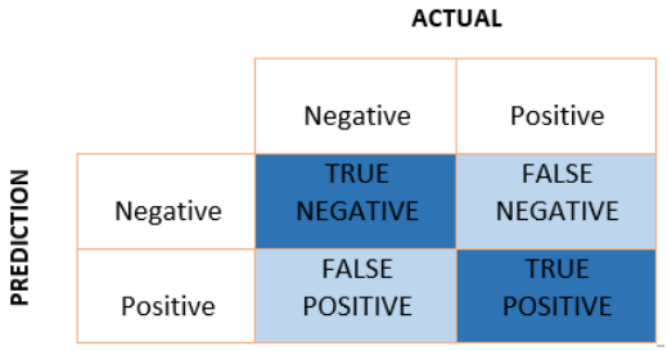


Fig 4.1 Confusion Matrix

Example: A machine learning model is trained to predict tumor in patients. The test dataset consists of 100 people.

Let suppose the model have confusion matrix like this:

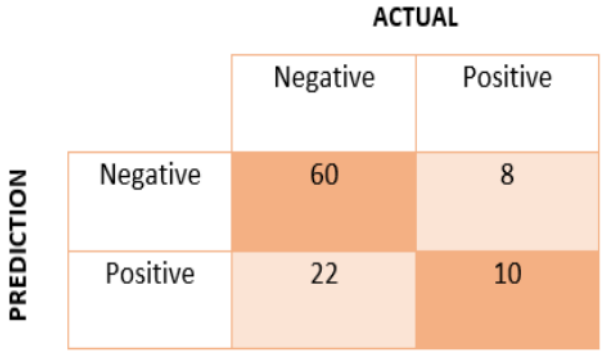


Fig 4.2 Example

True Positive (TP) — It gives information about how our trained model is correctly predicting the positive class (prediction and actual both taken positive in this case).

True Negative (TN) — The model correctly predicts the negative class (prediction and actual both are negative).For the example given above , there are **60** number of people who don't have tumors and are predicted as negative.

False Positive (FP) — The model gives the wrong prediction of the negative class (predicted-positive, actual-negative).For the example given above, there are **22 people** predicted as positive means they have tumour , although they don't have a tumor. FP is also called a **TYPE I** error.

False Negative (FN) — The model wrongly predicts the positive class (predicted-negative, actual-positive).For the example given above, there are **8 people** who actually have tumors but they are predicted negative means they are free from tumor. FN is also called a **TYPE II** error.

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

Accuracy: It represents how often is the classifier correct.
(TP+TN)/ total samples

Precision: Out of all the positives predicted, what percentage is truly positive.The precision value lies between 0 and 1.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Out of the total positive, what percentage are predicted positive. It is the same as TPR (true positive rate).

$$Recall = \frac{TP}{TP + FN}$$

F1 Score: Some quantity which can relate both precision and recall is needed. Since, precision and recall has inverse relation we can't simultaneously increase both quantities. So, we considered the harmonic mean of precision and recall which comes out to be known as F1 score. Both false positives and false negatives are considered for this. Therefore, it performs well on an imbalanced dataset.

Since, F1 score calculates harmonic mean it gives the same weight age to both precision and recall.

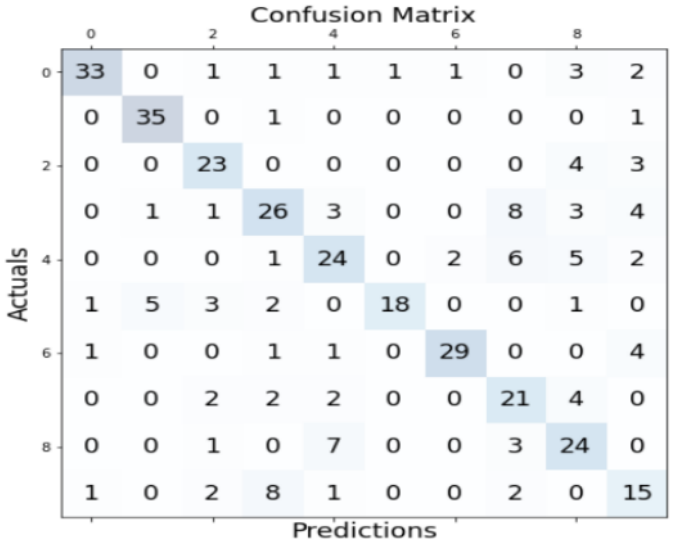
$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

4.2 Result

GTZAN dataset is used which contains 1000 songs belonging to 10 different genres and each genre contains 100 songs. After dividing the dataset into train test split a total of 357 songs is used for testing purpose which is also shown in support part of fig 4.3.Since, this project is based on classification Confusion matrix is used for evaluation purpose.

Below is the 10x10 confusion matrix where first row represents that out total 43 music files belonging to class 0 i.e, 'blues' 33 are predicted correct 1 file is predicted wrongly as 'country' 3 as 'reggae' 2 as 'rock' and so on. The principal diagonal of this matrix shows numbers of those test samples which are correctly predicted by the model. After analyzing all these final accuracy comes out to be 69%.

0:'blues', 1:'classical', 2:'country', 3:'disco' , 4: 'hip hop', 5:'jazz', 6:'metal', 7:'pop', 8:'reggae', 9:'rock'



	precision	recall	f1-score	support
1	0.92	0.77	0.84	43
2	0.85	0.95	0.90	37
3	0.70	0.77	0.73	30
4	0.62	0.57	0.59	46
5	0.62	0.60	0.61	40
6	0.95	0.60	0.73	30
7	0.91	0.81	0.85	36
8	0.53	0.68	0.59	31
9	0.55	0.69	0.61	35
10	0.48	0.52	0.50	29
accuracy			0.69	357
macro avg	0.71	0.69	0.69	357
weighted avg	0.72	0.69	0.70	357

Fig 4.3 Summary of K Nearest Neighbor ML algorithm:

The classification report of random forest algorithm is quite similar to KNN algorithm. In this a total of 352 music samples were taken for testing purpose belonging to 10 different classes. Since, the accuracy of this model is 48% which is lower than KNN model it can be seen it has lesser values in principal diagonal of its confusion matrix and almost 50% of samples belonging to particular class are wrongly predicted. For example ‘hip hop’ and ‘rock’ genre have least accuracy having more values as wrongly predicted than the actual correct predictions.

CLASSIFICATION REPORT

	precision	recall	f1-score	support
1	0.47	0.62	0.53	34
2	0.77	0.82	0.79	33
3	0.50	0.37	0.42	41
4	0.41	0.31	0.35	36
5	0.38	0.25	0.30	40
6	0.42	0.43	0.42	35
7	0.53	0.54	0.54	35
8	0.52	0.82	0.64	28
9	0.51	0.49	0.50	39
10	0.22	0.26	0.24	31
accuracy			0.48	352
macro avg	0.47	0.49	0.47	352
weighted avg	0.47	0.48	0.47	352

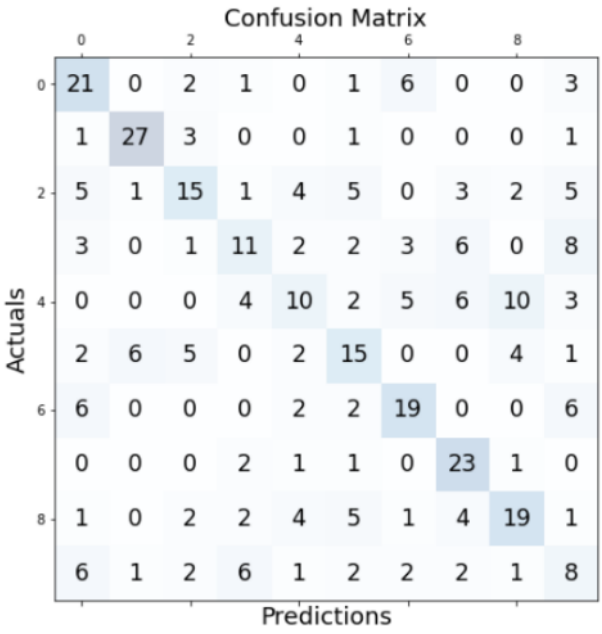


Fig 4.4 Summary of Random Forest Classifier ML algorithm:

Finally, the classification report of CNN algorithm is shown below.In this it can be seen that a total of 2497 samples are used in testing set. This has happened because for this model data augmentation is done means each 30 sec long 1000 music files is divided into 3 sec long audio file which implies now dataset contains 1000 music files for each genre generating a total of 10,000 music files. That's how after train test split we got a total of 2497 files.

Now, about confusion matrix observations are pretty much same the only difference is that accuracy has improved a lot after using CNN model which is 85%. Due to this, principal diagonal of confusion matrix contains most of the values for each genre which signifies during testing almost all the values are predicted as correct leaving very few which are giving wrong results. Class 6 (genre 'metal') is giving the best results out of 284 samples 249 are predicted correctly giving almost 90% f1 score and incorrectly predicted 16 'metal' samples as 'rock' which is the main source of error for this particular class. In some classes most of the wrong predictions are due to a particular another class while in some it is almost equally distributed.

CONFUSION MATRIX

	precision	recall	f1-score	support
0	0.88	0.82	0.85	261
1	0.97	0.91	0.94	242
2	0.83	0.78	0.80	288
3	0.77	0.86	0.81	222
4	0.87	0.85	0.86	226
5	0.88	0.90	0.89	283
6	0.95	0.88	0.91	284
7	0.87	0.86	0.86	237
8	0.83	0.88	0.86	233
9	0.66	0.75	0.70	221
accuracy			0.85	2497
macro avg	0.85	0.85	0.85	2497
weighted avg	0.85	0.85	0.85	2497

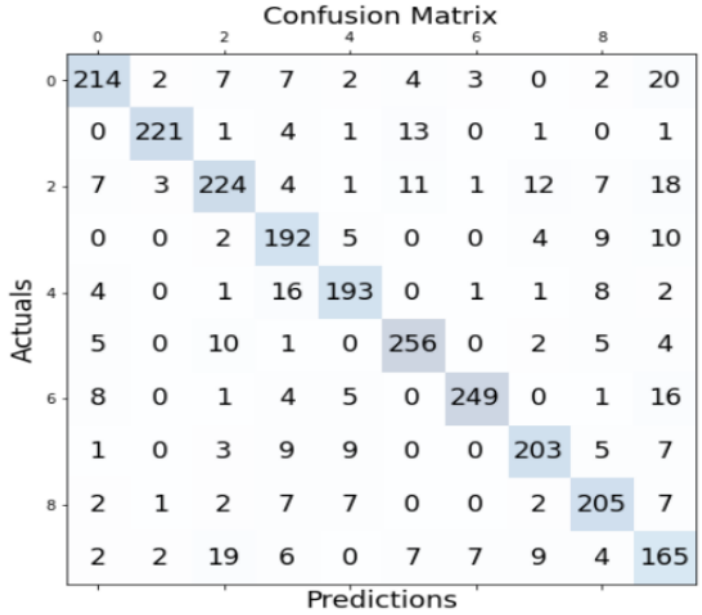


Fig 4.5 Summary of CNN Deep Learning algorithm:

Chapter : 5
Conclusion

5.1 Conclusion

In all models, frequency-based mel-spectrograms are used to produce high accuracy results. Whether it is only amplitude that provides information on the intensity, or how "loud noise", the frequency distribution provides information on the audio content. Additionally, mel-spectrograms are visible, and CNN works better with images. CNN did best, as we expected. It has taken too long to train again, but the increase in accuracy ensures that additional calculation costs. However, it was surprise to see the exact similarities between KNN, SVM, and neural transmission network.

1) It comes to notice that our CNN is too heavy for the 'rock' type. It results in least accuracy among all genres.
2) It has only managed to properly classify 70% of rock sound as rock, labeling others as mostly country or pop.
3) Additionally, it has misinterpreted country, disco and metal as rock music.
4) Rock has been a challenging genre - a test of the quality of rock mel-spectrogram means that many rock music quotes do not have easy beats of other genres such as hip-hop and disco.

5.2 Future Scope of Work

In the future, we hope to try other types of in-depth learning methods, as they have done very well. Given that this is a time series data, a specific type of RNN model may work well (GRU, LSTM, for example). We also want to know about the productive features of this project, which include some form of genetic modification (in the same way as production rival networks repaint images in Van Gogh style, but music in particular). Moreover, we suspect that we may have opportunities to pass on learning, for example in classifying music by artist or for ten years. As a future scope of our work we can do the following additions in our project.

1) As there are 1264 micro genres of music our project includes only 10 genres so to generalize it more genres can be used.
2) STFT is used to analyse the time-frequency domain represent but it does not give frequency component at particular instant of time so for better time and frequency resolution wavelet transform can be used.
3) Since, our dataset is a bit small we can combine different dataset to generate a big data which can cover as many variations of music as possible.
4) This work can be extended to speech recognition of different species.

References

[1] J.W. Picone, Signal modeling techniques in speech recognition. Proc. IEEE 81, 1215–1247 (1993)

[2] J. Volkmann, S. Stevens, E. Newman, A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. 8, 185–190 (1937)

[3] Z. Fang, Z. Guoliang, S. Zhanjiang, Comparison of different implementations of MFCC. J. Comput. Sci. Technol. 16, 582–589 (2000)

[4] G.K.T. Ganchev, N. Fakotakis, Comparative evaluation of various MFCC implementations on the speaker verification task, in Proceedings of International Conference on Speech and Computer (SPECOM) (2005), pp. 191–194

[5] J.S. Mason, X. Zhang, Velocity and acceleration features in speaker recognition, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (1991), pp. 3673–3676

[6] Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. “Long-term recurrent convolutional networks for visual recognition and description.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625- 2634. 2015.
.

[7] Denil M, Bazzani L, Larochelle H, et al. Learning where to attend with deep architectures for image tracking[J]. Neural computation, 2012, 24(8): 2151-2184.

[8] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in Neural Information Processing Systems. 2014: 2204-2212.

[9] Vinyals O, Kaiser , Koo T, et al. Grammar as a foreign language[C]//Advances in Neural Information Processing Systems. 2015: 2755-2763.

[10] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend[C]//Advances in Neural Information Processing Systems. 2015: 1684- 1692.

20185116

ORIGINALITY REPORT

17%	12%	2%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.jetir.org Internet Source	5%
2	Submitted to University of Birmingham Student Paper	5%
3	laptrinhx.com Internet Source	2%
4	www.sthda.com Internet Source	1%
5	www.coursehero.com Internet Source	1%
6	ijsrcseit.com Internet Source	1%
7	en.wikipedia.org Internet Source	1%

Exclude quotes	On	Exclude matches	< 50 words
Exclude bibliography	On		