DEEP LEARNING MODELS FOR MUSIC CLASSIFICATION

Thesis submitted to the
Indian Institute of Information Technology Guwahati
for award of the degree

of

Master of Technology

by

Tarun Kumar

under the supervision of

Dr.Moumita Roy

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

May 2024

CERTIFICATE

This is to certify that the thesis entitled **"Deep Learning Models for Music Classification"***, submitted by* **Tarun Kumar** *to the Indian Institute of Information Technology Guwahati, for the award of the degree ofMaster of Technology, is a record of bona fide research work carried out by him undermy supervision and guidance. The thesis, in my opinion, is worthy of consideration forthe award of the degree of Master of Technology in accordance with the regulations ofthe Institute. To the best of my/our knowledge, the results embodied in the thesis have not been submitted to any other university or institute for the award of any other degreeor diploma*

Dr.Moumita Roy,
Assistant Professor,
Department of Computer Science and Engineering
Indian Institute of Information Technology Guwhati

DECLARATION

I certify that

a. The work contained in this thesis is original and has been done by me under the general supervision of my supervisor(s).

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Tarun Kumar

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my guide and all anonymous referees from institute for their valuable comments which have helped to enhance the quality of thesis. Thanks for the support of my friends and mentors, who were with me while I am researching about this thesis. Their constant encouragement and guidance helped me complete my thesis without any difficulty.

I would also like to thank my mother, Pramila Devi, for her support and encouragement throughout my study.

ABSTRACT

In previous years the music genre classification is done using different methods. The methods used for the classification is deep learning and K neural networks. In Music Information Retrieval (MIR), research is going on for retrieving information from music and on building classifiers with better accuracy . Convolutional Neural Networks (CNN) and Support Vector Machines (SVM) are mostly used for music genre classification. We use a CNN approach for music genre classification using the GTZAN dataset. The proposed model shown that CNN model is effective for music genre classification and it also shown the importance of feature selection and parameter tuning in the performance of the model. Accuracy of the model is 70.90 percent.

Keywords: machine-learning, deep-learning, music genre classification, Information Retrievals , Music Genres, Convolution Neural Network (CNN).

Contents

Chapter 1
1. Introduction

Music is the universal language which is used to express feeling and the emotion beyond words. We can easily give a number of commonalities between music and language. For example, the musical score and text both are the symbolic transcript of their analog counterpart -music and speech, which manifest themselves by sounds organized in time. The fundamental difference that distinguishes music from every other languages is that the essence of music exists mainly in its audio performance .Humans can comprehend the meaning of language by reading text. But, we cannot fully understand the meaning of a music by merely reading its musical score.

Genre is one of the most common of factors distinguishing music pieces. Music genre classification, a subset of music information retrieval, is a challenging and progressive task in the artificial intelligence domain. It involves machine learning concepts and algorithms to recognize the genre of a particular music audio file, the style or category of the music. For example the algorithm tries to differentiate between a rock music file and a classical music file based on the features of the audio.

Among the many commercial applications of Deep Learning, Music Signal Processing has received an increasing amount of attention over the last decade. This work reviews the most recent developments of Deep Learning in Music signal processing. Two main applications - Music Information Retrieval and Music Generation. The broad field of DL in music-related applications can be called Music Deep Learning (MDL) and can be divided into two categories, Music Information Retrieval (MIR) [7] and Music Generation (MG) . MIR refers to the extraction of characterizing information from music data. These information can be used for wide range of applications -genre classification [1], [2], music recommendation [8], music source separation , singing voice detection, instrument recognition, music emotion recognition and transcription . All of the above applications aid in the digital preservation of music, by constructing and managing song databases, as well as the study of different music genres.

Music genre classification is automatically identifying the genre of a given audio signal. It has many applications in the music industry, including recommendation systems, personalized playlists, and content-based music retrieval. Traditional methods for music genre classification include feature extraction and classification using algorithms k-nearest neighbor, decision trees, and support vector machines. But, these methods require handcrafted features and it is suffering from low accuracy.

Convolutional Neural Networks (CNN) played very important role in the field of music genre classification as it automatically extract features from raw audio signals. CNNs model can learn high level representations of audio signals by performing convolution operations on the audio waveform. Support Vector Machines (SVM) is also widely used for music genre classification due to their ability to handle high-dimensional data and achieve high accuracy. Random forest combines the output of multiple decisions trees to reach a single result.
Music genre classification is a fundamental problem in the field of music information retrieval. We propose a CNN approach for music genre classification using GTZAN Dataset.

The GTZAN dataset is the benchmark dataset which is widely used for music genre classification, and it consists of 1000 audio tracks which is equally distributed across 10 different music genres. The 10 music genres- blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each music file is 30 seconds long and it is sampled at 22,050 Hz, and stored in the .WAV format. The goal of the dataset is to develop accurate and robust machine learning models which can classify the genre of the given music track accurately. The GTZAN dataset has been used as a standard benchmark dataset for evaluating the performance of various music genre classification algorithms and has played a significant role in advancing the field of music information retrieval.

The use of CNNs in music genre classification has opened new avenues for the development of more accurate and efficient classification methods. Using the large datasets like GTZAN, researchers are expected to continue exploring new techniques and approaches for music genre classification for improving the quality of music recommendation and discovery systems

We experienced that Spotify and YouTube create playlists and suggest the next song based on our previous listening patterns. These patterns are created based on the tone, mood, or genre of songs which we listen to. Spotify as we use in our daily life for listening the songs and knows that it maintain very large databases of music and uses music genre classification. Hence, classifying music based on genres suggests the next songs to a listener, design the playlists for new recommendations, or filter unwanted content.

1.1 Problem Statement

With the growth of online music databases and easy access to music content, people find it increasing hard to manage the songs that they listen to. One way to categorize and organize songs is based on the genre, which is identified by some characteristics of the music such as rhythmic structure, harmonic content and instrumentation [5].

Music genre classification is an important task in the field of music information retrieval. It involves the automatic identification and classification of music into different music genres based on their audio features. Music genre classification has several real-world applications such as music recommendation, content-based music retrieval, and personalized music services. But the task of music genre classification is challenging due to the subjective nature of music and the complexity of audio signals. Traditional methods for music genre classification has low accuracy and limits their practical applications. Hence there is a need for more accurate and robust methods for music genre classification that can handle the diversity and complexity of music. The problem statement for this research is to propose a CNN approach for music genre classification that can achieve high accuracy on the GTZAN Dataset. There are several problem statements in music genre classification in machine learning that researchers have been addressing.

The ultimate goal of the researchers is to develop accurate and efficient music genre classification models which can be used in different music-related applications such as music recommendation systems and personalized playlists.

Some of the important points in the music genre classification which can be seen as problems are as follows:

1. Classification of music genre in practive: Practical that is real time classification of audio signals for music genre classification need fast and efficient machine learning methods that can process the audio data in real- time.

2. Similarity between Inter-genre: Some music genres may have similar acoustic features, and so it makes difficult for a machine learning model to distinguish between them.

3. Dimensional of audio Data: Audio signals are high-dimensional and so require large amounts of storage and processing power and which can be a challenge for some machine learning methods.

4. Feature Extraction of audio data: Traditional music genre classification methods need manual extracting of features from audio signals, which is a time-consuming and complex

process. The features extracted may not always be effective in accurately representing the audio signal and may limit the accuracy of the classification model.

5. Shortage of data and labelled data: There is a shortage of labelled audio data for certain music genres, which can make it difficult to train accurate machine learning models for these genres.

We use new machine learning methods, deep learning models and the use of large labelled datasets to avoid these problems.

1.2 Objectives

The objectives of the research are to do a thorough survey on music classification machine-learning-based and deep neural architecture. We will prepare music classification dataset to develop music genre classification model using Convolutional Neural Networks (CNN) for the Kaggle GTZAN Dataset Music Genre Classification.

We will propose a Convolutional Neural Network (CNN) Based Deep Learning System to classify the music genre.

We are working for the following objectives:

1. Study the literature on machine learning and deep learning CNN model for music genre classification.

2. Preprocessing GTZAN Dataset for feature extraction and classification.

3. Design and implement CNN model approach for music genre classification.

4. Assess the accuracy of the proposed model for music genre classification.

1.3 Background

Deep Learning has strong computational tools, which have been extensively used in data and signal processing, with promising results. Among the many commercial applications of Deep Learning, music signal processing has received an increasing amount of attention over the last decade. Content are of various kinds: images, text and music and music is focus of our analysis. Deep Learning in music-related applications is called Music Deep Learning (MDL) and can be divided into two categories- music information retrieval (MIR) and music generation (MG). Music classification comes under music information retrieval (MIR) category.

Music Information Retrieval (MIR)

MIR refers to the extraction of characterizing information from music data. Characterizing information from music data can be used for many applications,such as genre classification ,music recommendation ,music source separation , singing voice detection ,instrument recognition, music emotion recognition, Instrument/voice identification and transcription. All these applications helps in the digital preservation of music, by constructing and managing song databases, as well as the study of different music genres.

Automatically determining the genre of a music tracks using computational methods is known as music genre classification. It is a critical task in the field of music information retrieval that requires the creation of algorithms and techniques for organizing, searching, and retrieving music data.

Music genres are categories used to categorize music based on musical and cultural characteristics. For many years, various approaches to music genre classification have been proposed. Traditional approaches use handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Centroid, and Zero Crossing Rate. These features are then fed into classification machine learning algorithms such as k-nearest neighbours, decision trees, and support vector machines.

In the last few years, deep learning methods are used in music genre classification including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have been proposed. And these models outperform traditional methods and improve classification accuracy significantly.

Music genre classification has several applications, including music recommendation systems, music search engines, and music licensing and copyright management. It is also an important task for musicologists and music enthusiasts, as it provides a systematic way of organizing and analyzing large collections of music.

1.4 Challenges and Research Questions

We see below as the challenges and research questions :-

1. Can the performance of the proposed model will be better than the traditional methods ?

2. Can we use CNN model for music genre classification on the GTZAN Dataset?

3. Does different hyperparameters affect the performance in the proposed model?

4. Can the additional features - lyrics and metadata improve the accuracy of the proposed model?

5. Can we apply CNN model for the larger datasets with the same accuracy?

1.5 Motivation

In machine learning or deep learning models, main attention is given to computer vision or natural language sub-domain problems. But in today's date, demand to process audio data is increasing with emerging advancements in technologies like Google Home and Alexa that extract information from voice signals. Now a days working with audio data has become a new trend and area of study. But from the literature survey, we found that current model does not perform well for timbre and audio quality. The possible applications are voice recognition, music classification, tagging, and generation, and are paving the way for audio use cases to become the new era of deep learning. Mostly for this project, concentration is basically given on automatic music classification using audio /signal. The motivation is to use now widely available corpus to learn musical styles automatically and to classify music genre based on this. As more large scale musical datasets are made available, a machine learning-based music classification system will be able to automatically classify musical style from these datasets.

There is a growing interest by some big tech companies in this field like Google in June 2016 of the Magenta research project and the creation by Spotify in September 2017 of the Creator Technology Research Lab. And finally it will blur the line between music creation and music consumption through the personalization of musical content.

Chapter 2

2. Literature Review

This section includes all the techniques and approaches that would to help the new challenges of Music classification. The growing proximity of online data and review is a blossom for Music classification Analysis and another research area. New technologies are most welcome to overcome the challenges as well as understand people’ s opinions.

2.1 Music classification

Costa Y.M.G , Oliveira L. S., and Silla Jr .C.N, [1] propose:

2.1.1 Introduction

In the above paper, author compare the results obtained with a Convolutional Neural Network (CNN) with the results obtained by using handcrafted features and SVM classifiers. Author have performed experiments fusing the results obtained with learned features and handcrafted features to assess the complementarity between these representations for the music classification task.

2.1.2 Proposed methodology

CNN model is used. Experiment showsthat the CNN compares favourably to other classifiers in several scenarios, hence, it is a very interesting alternative for music genre recognition. If we consider African database, the CNN surpassed the handcrafted representations and also the state-of-the-art by a margin

2.1.3 Conclusion

Evaluated the CNN model for music content characterization using three music databases with distinct characteristics. Experiments have shown that the CNN compares favourably to other classifiers in several scenarios. Combination of a CNN with an SVM trained with Robust Local Binary Pattern (RLBP) achieved 92% of recognition rate, which is to the best of our knowledge, the best result (using the artist filter) on this dataset so far.

Senac C., Pellegrini T., Mouret F., and Pinquier J [2] propose :

2.2.1 Introduction

Music Genre Classification (MGC) is an active research topic in the field of Music Information Retrieval (MIR) as it is one of the most common ways to manage digital music databases. In most systems, MGC consists of extracting a set of features from the raw audio signal, optionally performing feature selection, and making a classification based on machine learning methods. Several works have been based on the extraction of discriminating audio features categorized as frame-level, segment-level or song level features. Frame-level features, such as Spectral Centroid, Spectral Roll off, Octave-based Spectral Contrast, Mel Frequency Cepstral

Coefficients, describe the local spectral characteristics of audio signal and are extracted from short time windows (or frames) during which the signal is assumed to be stationary.

2.2.2 Proposed methodology

In this paper CNN Model is used and CNNs is trained in such a way that filter dimensions are interpretable in time and frequency, results show that only eight music features are more efficient than 513 frequency bins of a spectrogram and that late score fusion between systems based on both feature types reaches 91% accuracy on the GTZAN database.

2.2.3 Conclusion

Proposed model CNN use a map of eight musical features as inputs of a CNN. Results show the relevance of our eight music features: global accuracy of 89.6% against 87.8% for 513 frequency bins of a spectrogram. The late score fusion between systems based on both feature types reaches 91% accuracy on the GTZAN database. For future work, it is planned to make an early fusion of the two networks in order to have a global classifier. We also have to test our method with other databases with distinct characteristics such as "The Latin American Music database" or ethnic music

Hirvonen T [3] proposes :-

2.3.1 Introduction

In the above paper, author use of Convolutional Neural Networks for spatial audio classification. In compared to traditional methods that use hand-crafted features and algorithms, author show that a CNN in combination with generic preprocessing will give good results and allows for specialization to challenging conditions.

2.3.2 Proposed methodology

We use a total of seven layers with learned parameters, and a final softmax layer for classification. The first four layers have convolutional connectivity. A seemingly useful strategy for lowering the number of network parameters was to use short strides for the convolutional layers. In this study, the stride was 4, with an overlap of 2. Pooling layers typical to vision research are omitted here. They do not seem useful for audio, as positional pattern shifts do not occur in the whitened spectrograms as in visual images. The last three layers have full connectivity. Use standard backpropagation with stochastic gradient descent to train the CNN. The learning hyperparameters were: learning rate = 0.01 and momentum = 0:9. Throughout the training, learning reate was decreased sequentially by multiplying it by 0.1 every 160000 training samples. The batch size for the weight up- dates in training was 16 training samples. It has the accuracy of 94% .

2.3.3 Conclusion

In this paper, experiments detailing the use of CNNs to spatial audio analysis were presented. This represents a novel application of a previously established supervised learning method. An important aspect of the paper is to corroborate the use of an audio pre-processing technique that allows for a good classification performance in various even seemingly unrelated tasks.

Chapter 3
3. Methodology

3.1 Dataset description

Data is the most valuable resource for the machine-generated model. “Data is the new oil.”

Below are the datasets available to use for music classification tasks. In this project we use **Kaggle GTZAN Genre Collection** dataset.

- GTZAN Genre Collection
  https://www.kaggle.com/datasets/carlthome/gtzan-genre-collection

- MusicNet Dataset
  https://zenodo.org/record/5120004#.YXDPwKBlBpQ

- Groove MIDI Dataset (GMD)
  https://magenta.tensorflow.org/datasets/groove

- JSB-Chorales-dataset
  https://github.com/czhuang/JSB-Chorales-dataset
  https://www.kaggle.com/datasets/nadiacarvalho/jsb-chorales-signallike-embeddings

- **GTZAN Genre Collection**

This dataset was used for the well known paper in genre classification "Musical genre classification of audio signals" by G. Tzanetakis and P. Cook in IEEE Transactions on Audio and Speech Processing 2002.

The dataset has the following folders:

- Genres original — A collection of 10 genres with 100 audio files each, all having a length of 30 seconds (the famous GTZAN dataset, the MNIST of sounds)

- Images original — A visual representation for each audio file. One way to classify data is through neural networks because NN's usually take in some sort of image representation.

- 2 CSV files — Containing features of the audio files. One file has for each song (30 seconds long) a mean and variance computed over multiple features that can be extracted from an audio file. The other file has the same structure, but the songs are split before into 3 seconds audio files.

The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format.

The genres are:

- blues
- classical
- country
- disco
- hiphop
- jazz
- metal
- pop
- reggae
- rock

- **MusicNet Dataset**

  MusicNet is a collection of 330 freely-licensed classical music recordings, together with over 1 million labels indicating the precise time of each note in every recording, the instrument that plays each note, and the note's position in the metrical structure of the composition.

  https://zenodo.org/record/5120004#.YXDPwKBlBpQ
  This repository consists of 3 top-level files:
  musicnet.tar.gz - This file contains the MusicNet dataset itself, consisting of PCM-encoded audio wave files (.wav) and corresponding CSV-encoded note label files (.csv). The data is organized according to the train/test split described and used in "Invariances and Data Augmentation for Supervised Music Transcription".

  musicnet_metadata.csv - This file contains track-level information about recordings contained in MusicNet. The data and label files are named with MusicNet ids, which you can use to cross-index the data and labels with this metadata file.

  musicnet_midis.tar.gz - This file contains the reference MIDI files used to construct the MusicNet labels.

- Groove MIDI Dataset (GMD)

  The dataset is made available by Google LLC. The dataset contains about 13.6 hours, 1,150 MIDI files, and over 22,000 measures of drumming.
  GMD is available as a zip file containing the MIDI and WAV files as well as metadata in CSV format.
  The metadata file (info.csv) has the following fields for every MIDI/WAV pair:

Field	Description
drummer	An anonymous string ID for the drummer of the performance.
session	A string ID for the recording session (unique per drummer).
id	A unique string ID for the performance.
style	A string style for the performance formatted as "<primary>/<secondary>". The primary style comes from the Genre List below.
bpm	An integer tempo in beats per minute for the performance.
beat_type	Either "beat" or "fill"
time_signature	The time signature for the performance formatted as "<numerator>-<denominator>".
midi_filename	Relative path to the MIDI file.
audio_filename	Relative path to the WAV file (if present).
duration	The float duration in seconds (of the MIDI).
split	The predefined split the performance is a part of. One of "train", "validation", or "test".

- JSB-Chorales-dataset

  JSB Chorales prepared with train/test separation.
  The **JSB** chorales are a set of short, four-voice pieces of music well-noted for their stylistic homogeneity. The chorales were originally composed by Johann Sebastian Bach in the 18th century. He wrote them by first taking pre-existing melodies from contemporary Lutheran hymns and then harmonising them to create the parts for the remaining three voices. The version of the dataset used canonically in representation learning contexts consists of 382 such chorales, with a train/validation/test split of 229, 76 and 77 samples respectively.

3.2 Data Preprocessing

Preparing Data

We predict a music genre from a randomly chosen 30-second music audio track. This means that the dependent variable (the one we are attempting to measure) is the song genre.

We convert the audio signals to spectrograms and dividing them into 128x128 pixel images. The we divide the available data into training, validation, and testing and it is critical to the training and evaluation of a machine-learning model performance in a robust and unbiased manner. It will help to assess the model's ability to generalise to unseen data similar to the real-world scenario.

The images' pixel values were normalised between 0 and 1 by dividing the RGB value by 255.0 as an act of preprocessing. This was essential for the models to interpret the dataset.

We check for null or missing values in the dataset. As we check and found that there is a missing values in the kaggle GTZAN dataset and the missing values or corrupted file is in the jazz dataset. And after verifying we found that the file 'jazz.00054.wav' is corrupted in the 'genres_original' directory and it means that no spectrogram was generated for it in the 'images_original' directory. To solve this problem and avoid potential bias, we will trim down all genre datasets to only 99 files in the training-validation-test splitting process. I have chosen an 80:9:10 split for my training, validation and testing data respectively.

3.3 Research Methodologies

Deep Learning Methodologies

Music classification algorithms can be generally categorized as follows: Deep Learning: LSTM, CNN, VAE, GAN etc., Evolutionary Computation, EC.

Deep Learning methodologies are Convolutional Neural Networks, Recursive Neural Networks, Generative Adversarial Nets, Variational Auto-Encoder, Transformer, Evolutionary Algorithms.

Audio File Overview

Audio i.e. sounds are pressure waves and can be represented by numbers over a time period. Audio files are generally stored in .wav format and so need to be digitized, using the concept of sampling.

Sample rate is the number of samples (data points) per second in a sound. For example: if the sampling frequency is 44 khz, a recording with a duration of 60 seconds will contain 2,646,000 samples. In practice, sampling even higher than 10x helps measure the amplitude correctly in the time domain.

3.4 Proposed methodology : Convolutional Neural Network (CNN)

When we say Convolution Neural Network (CNN), generally we refer to a 2 dimensional CNN which is used for image classification. However there are two other types of Convolution Neural Networks used in the real world, which are 1 dimensional and 3-dimensional Convolution Neural Network.
1D Convolutional Neural Networks are similar to well-known and more established 2D Convolutional Neural Networks. 1D Convolutional Neural Networks are used mainly used on text and 1D signals.

Convolutional Neural Network (CNN) is a type of neural network architecture that is particularly effective in finding hidden patterns. 1D CNN can be used for processing sequential data, such as time series data, in our case daily weather data. CNN had two main layers

1) Feature extraction layer convolution and subsampling layers.
2) Fully connected layers this layer use for classification and prediction.

In CNN, the convolution layers have various parameters, such as feature maps, filters sizes, activation functions, stride, and padding. The output from the kernel set sends to the next layer, which in turn will form a convolution layer. Convolution layer is the starting layer of CNN architecture, which consists of height, width and thickness in case of 2D convolution and length and thickness in case of 1D convolution. This architecture consists of many convolution layers, with the output of the previous layer used as input features to the next layer. The convolution layer is used to extract the hidden features and extracted features then feed to fully connected layers for classification or prediction.

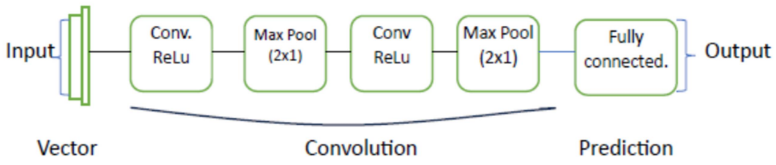

Fig 1

Output of convolution layers is calculated as

$$= \frac{N-F+2P}{s} + 1$$

Where N is input width, F is the filter width (Kernal), P is padding, S is the stride.
ReLu Activation $(x) = Max(0_t, X)$

The output of the convolution layer enters the pooling layer. This pooling layer performs down sampling. Max pooling speeds up computation and control overfitting. There are different types of pooling, Max Pooling select max value among two values if pooling layer is (1X2).

The last layer is used for prediction, which is composed of several layers, and each layer is composed of fully connected layer (Multilayer Perceptron). This layer gets input as extracted feature from convolution layers and output as classification or prediction value as required. Above we have showed an example of MLP architecture consisting of three layer namely the input layer, hidden layer and output layer. At the input layer, the input varied with Xn. Om the hidden layer, there are weights (Vij) and bias (Voj) and Z as hidden layer. At output layer too there are weights (Wij) and bias (Woj) with the output data varied with Y. At the input layer, there is no computational process, but at the input layer, the Xi input signal sends to the hidden layer. In the hidden layer and output layer, the computation process of weight and bias occurs, and the magnitude of the output of the hidden layer and output layer is calculated based on the activation function.

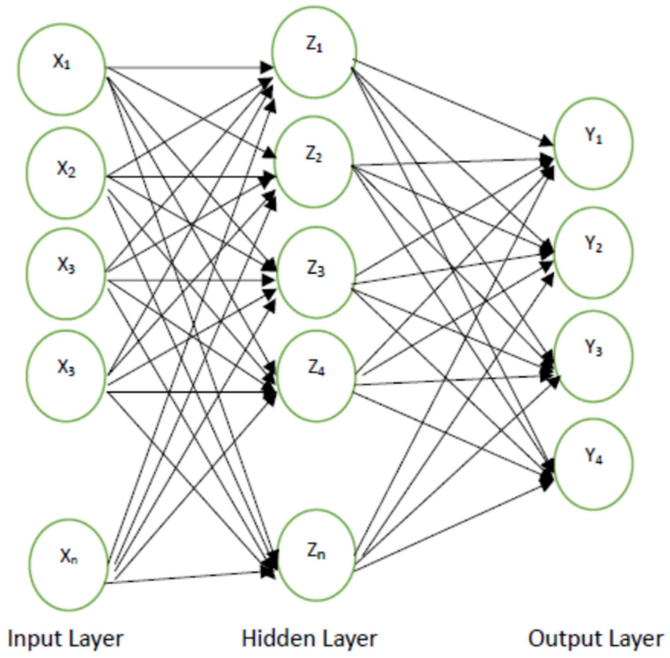

Fig 2

Sound are pressure waves, and these waves can be represented by numbers over a time period. These air pressure differences communicates with the brain. Audio files are generally stored in .wav format and need to be digitized, using the concept of sampling.
We load and visualize an audio file in python. Librosa is a Python library that helps us work with audio data. We can Play Audio using IPython.display.Audio, we can play the audio file in a Jupyter.

Waveform visualization : Using Matplotlib and Librosa we visualize the sampled signal and plot it.
Spectrogram : A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. They are time-frequency portraits of signals. Using a spectrogram, we can see how energy levels (dB) vary over time.
Normalization : A technique used to adjust the volume of audio files to a standard set level; if this isn't done, the volume can differ greatly from word to word, and the file can end up unable to be processed clearly.

The 1D Convolution Neural Network learn a representation and a discriminant directly from the raw audio signal. Several convolutional layers capture the time-frequency characteristics of the audio signal and learn various filters relevant to the music genre recognition task.

The CNN model has five convolutional layers , ReLU activation functions and max pooling layers. Training and Testing: The model is trained using the Adam optimizer which has a learning rate of 0.0001 and a batch size of 32. The model was trained for 100 epochs, with early stopping used to prevent overfitting. The accuracy of the model was evaluated using the classification accuracy, which is the percentage of correctly classified samples. The testing data was used to evaluate the performance of the proposed model.
The hyperparameters is tuned by using a grid search approach. The number of filters, kernel size, and dropout rate are the hyperparameters tuned for the CNN model.

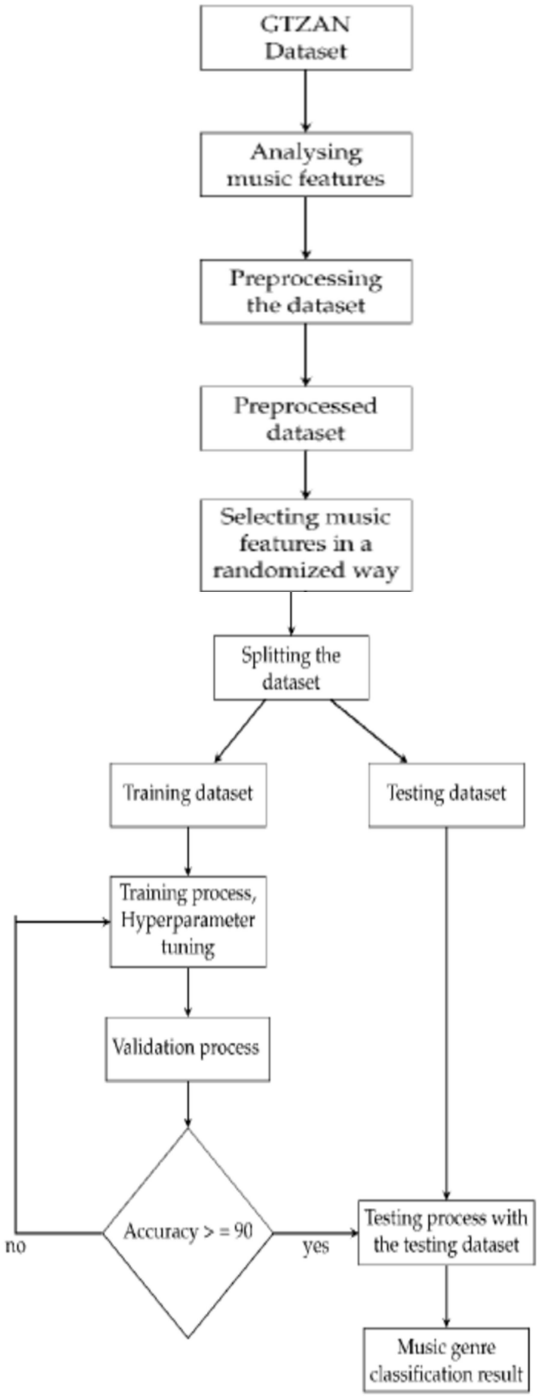
GTZAN Dataset
Analysing music features
Preprocessing the dataset
Preprocessed dataset
Selecting music features in a randomized way
Splitting the dataset
Training dataset
Testing dataset
Training process, Hyperparameter tuning
Validation process
Accuracy > = 90
no
yes
Testing process with the testing dataset
Music genre classification result

Fig 3

4. Analysis and Evaluation

Sound are pressure waves, and these waves can be represented by numbers over a time period. These air pressure differences communicates with the brain. Audio files are generally stored in .wav format and need to be digitized, using the concept of sampling.

We load and visualize an audio file in python. Librosa is a Python library that helps us work with audio data. We can Play Audio using IPython.display.Audio, we can play the audio file in a Jupyter.

Waveform visualization : Using Matplotlib and Librosa we visualize the sampled signal and plot it.

Spectrogram : A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. They are time-frequency portraits of signals. Using a spectrogram, we can see how energy levels (dB) vary over time.

Normalization : A technique used to adjust the volume of audio files to a standard set level; if this isn't done, the volume can differ greatly from word to word, and the file can end up unable to be processed clearly.

1. Plot Raw Wave Files

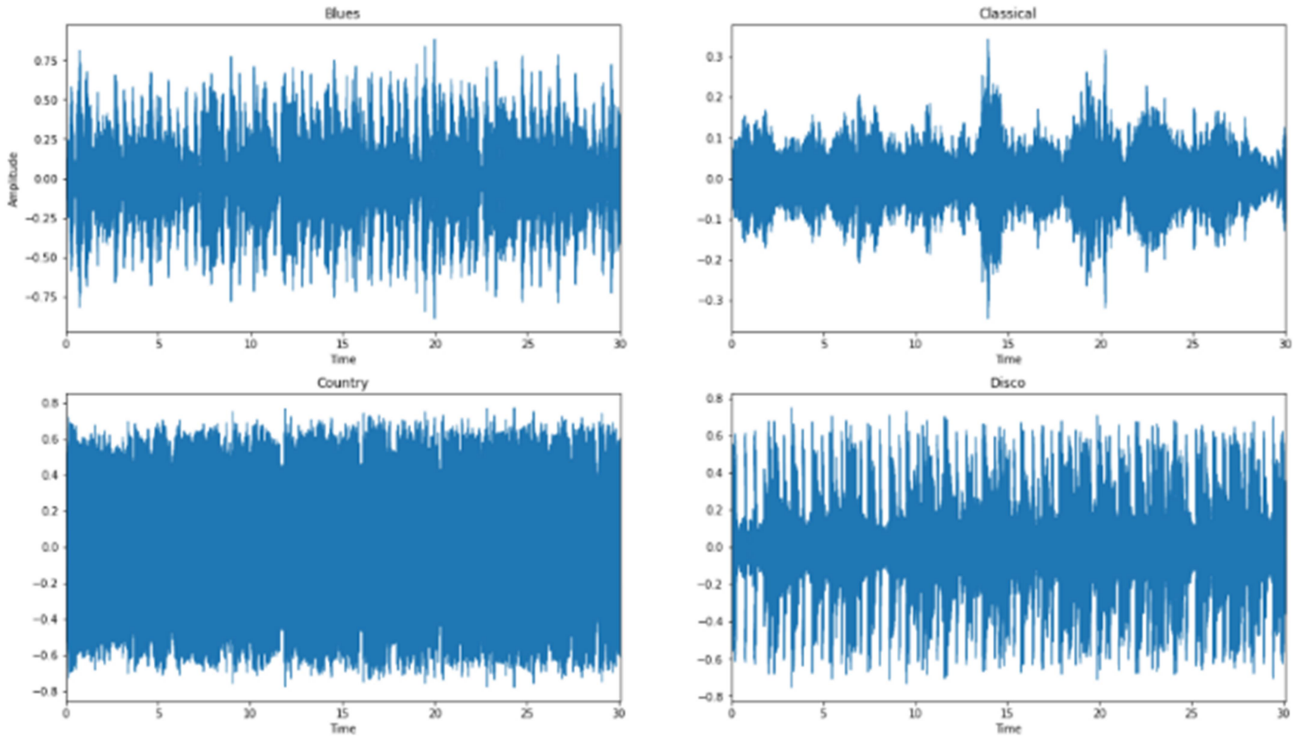

Fig 4

Waveforms are visual representations of sound as time on the x-axis and amplitude on the y-axis. We use it for read the audio data and visually compare and contrast which genres looks more similar than others.

```
plt.figure(figsize=(14, 5))
#plotting the sampled signal
# Visualizing Audio File as a waveform
librosa.display.waveshow(data, sr=sample_rate)
```

```
<librosa.display.AdaptiveWaveplot at 0x7e15a046ead0>
```

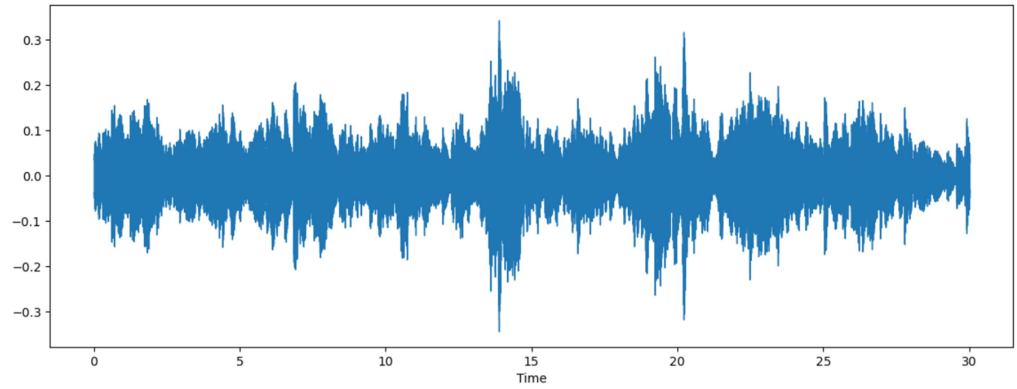

Fig 5

2. Spectrograms

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. They are time-frequency portraits of signals. Using a spectrogram, we can see how energy levels (dB) vary over time.. In 3-dimensional plot, it looks like waterfalls but in 2-dimensional arrays, the first axis is frequency and the second axis is time.

```
#x: numpy array
X = librosa.stft(x)
#converting into energy levels(dB)
Xdb = librosa.amplitude_to_db(abs(X))

plt.figure(figsize=(20, 5))
librosa.display.specshow(Xdb, sr=sr, x_axis='time', y_axis='hz')
plt.colorbar()
```

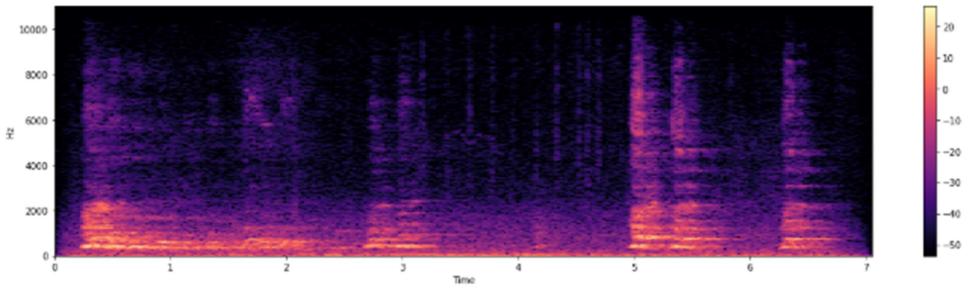

Fig 6

```
[37]: stft = librosa.stft(data)
      stft_db = librosa.amplitude_to_db(abs(stft))
      plt.figure(figsize=(14, 6))
      librosa.display.specshow(stft_db, sr=sr, x_axis='time', y_axis='hz')
      plt.colorbar()

[37]: <matplotlib.colorbar.Colorbar at 0x7fa44f047160>
```

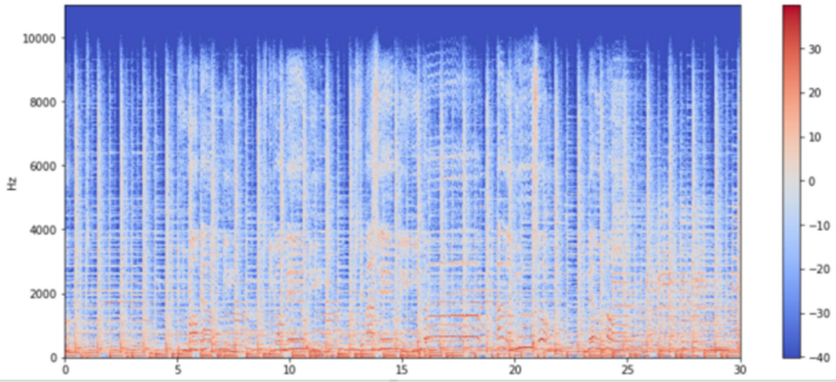

Fig 7

The vertical axis shows frequencies (from 0 to 10kHz) and the horizontal axis shows the time of the clip.

Scaling the Features :

We use Standard scaler to standardize features by removing the mean and scaling to unit variance.

The standard score of sample x is calculated as:

z = (x - u) / s

Dataset Standardization is a common requirement in machine learning because: they behave badly if the individual features do not more or less look like standard normally distributed data.

```
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
train_features = scaler.fit_transform(train_features.reshape(-1, train_features.shape[-1])).reshape(train_features.shape)
```

we use epoch = 100 for training model. We use Adam optimizer in CNN for training the model. All hidden layers are using the RELU activation function and the output layer uses the softmax function. The loss is calculated using the sparse_categorical_crossentropy function.

We use Dropout for avoiding the overfitting. We use the Adam optimizer as it provide us the best results after evaluating other optimizers.

We can increase the model accuracy by further increasing the epochs but after a certain period, we may achieve a threshold, so the value should be changed accordingly.

We train the model on a training set, evaluate its performance on a validation set, and then report the model's accuracy, precision, recall, and F1 score in case of CNN model for music genre classification.

Experiments and outputs:

First we will check if music audio file is in given dataset or not :

```
    genre = classifier.predict(scaled_row)

    # Print the predicted genre
    print(f"Predicted genre for {audio_file}: {genre[0]}")
    # Example usage of the find_genre function
find_genre('blues.00000.wav')
```

```
Error: blues.00000.wav not found in dataset.
```

Using below we find the Categories of Music Genres:

```
# Predict the genre of the test set
y_pred = classifier.predict(X_test)
```

```
['blues' 'classical' 'country' 'disco' 'hiphop' 'jazz' 'metal' 'pop'
 'reggae' 'rock']
```

Predicting the music genre of the given audio file :

```
    # Print the predicted genre
    print(f"Predicted genre for {audio_file}: {genre[0]}")
    # Example usage of the find_genre function
find_genre('blues.00000.0.wav')
```

```
Predicted genre for blues.00000.0.wav: blues
```

CNN Model :

```
def cnnmodel_tuned(input_shape):
    clear_session()
    model = Sequential()
    model.add(Conv2D(filters=32, kernel_size=(3, 3), strides=(1, 1), input_shape=X_train.shape[1:]))
    model.add(BatchNormalization())
    model.add(LeakyReLU())
    model.add(MaxPool2D(pool_size=(2, 2), strides=(2, 2)))
    model.add(Conv2D(filters=64, kernel_size=(3, 3), strides=(1, 1)))
    model.add(LeakyReLU())
    model.add(MaxPool2D(pool_size=(2, 2), strides=(2, 2)))
    model.add(Dropout(0.25))
    model.add(Conv2D(filters=128, kernel_size=(3, 3), strides=(1, 1)))
    model.add(LeakyReLU())
    model.add(AveragePooling2D(pool_size=(2, 2), strides=(2, 2)))
    model.add(Dropout(0.25))
    model.add(Conv2D(filters=256, kernel_size=(3, 3), strides=(1, 1)))
    model.add(LeakyReLU())
    model.add(AveragePooling2D(pool_size=(2, 2), strides=(2, 2)))
    model.add(Conv2D(filters=512, kernel_size=(3, 3), strides=(1, 1)))
    model.add(LeakyReLU())
    model.add(GlobalAveragePooling2D())
    model.add(Dense(10, activation='softmax'))  # Assuming 10 classes
    return model
```

The accuracy achieved for the CNN model used in the project is **70.90 percent** for GTZAN Genre Classification dataset.

5. Conclusion

Our experimental finding shows that a convolutional neural network (CNN) model performs better on Music Classification. This project we talked about the importance of feature selection and parameter tuning to get high accuracy in music genre classification. We carefully selected the most relevant features for genre classification and optimized the parameters to get the better performance. And we can use it for other related tasks such as mood detection or artist identification. The results demonstrate the potential of deep learning and machine learning techniques in analyzing and classifying music, which can have various applications in the music industry and beyond.

6. Future Work

Exploring various feature selection and parameter tuning can improve the performance of the model and provides into the most relevant features and parameters for music genre classification. We can use more advanced deep learning techniques such as Recurrent Neural Networks (RNNs) and Transformer Networks, which have provided better results in music information retrieval tasks. We can also try to use transfer learning, in which pre-trained models such as VGG or ResNet, are used as a starting point and fine-tuned for the specific task of music genre classification.

References

[1] Costa Y.M.G , Oliveira L. S., and Silla Jr .C.N, An evaluation of convolutional neural networks for music classification using spectrograms, Applied Soft Computing., vol. 52, pp. 28–38, Mar. 2017.

[2] Senac C., Pellegrini T., Mouret F., and Pinquier J., Music feature maps with convolutional neural networks for music genre classification, in Proc. 15th Int.Workshop Content-Based Multimedia Indexing, Jun. 2017, pp. 1–5.

[3] Hirvonen T., Classification of spatial audio location and content using convolutional neural networks, in Proc. 138th Audio Eng. Soc. Conv., May 2015, pp. 1–10.

[4] Hershey S., Chaudhuri S., Ellis D. P. W., Gemmeke J. F., Jansen A., Moore R. C., Plakal M., Platt D., Saurous R. A., Seybold B., Slaney M., Weiss R. J., and Wilson K., CNN architectures for large-scale audio classification, arXiv:1609.09430v2 [cs.SD] 10 Jan 2017.

[5] Hareesh Bahuleyan, Music Genre Classification using Machine Learning Techniques, arXiv:1804.01149v1 [cs.SD] 3 Apr 2018

[6] Wyse L., Audio spectrogram representations for processing with Convolutional Neural Networks, arXiv:1706.09559v1 [cs.SD] 29 Jun 2017

[7] Purwins H., Li B., Virtanen T., Schlüter J., Chang S.-Y., and Sainath T., Deep learning for audio signal processing, arXiv:1905.00078v2 [cs.SD] 25 May 2019

[8] Lu M., Pengcheng D., and Yanfeng S., Digital music recommendation technology for music teaching based on deep learning, Wireless Commun. Mobile Comput., vol. 2022, pp. 1–8, May 2022.

[9] Choi, K., Fazekas, G., & Sandler, M. (2016), Automatic tagging using deep convolutional neural networks. In Proceedings of the 17th International Society for Music Information Retrieval Conference, New York, USA, arXiv:1606.00298v1 [cs.SD] Jun. 2016.

[10] Ndou N., Ajoodha R., and Jadhav A., Music genre classification: A review of deep-learning and traditional machine-learning approaches, in Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS), Apr. 2021, pp. 1–6.