CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

# II Trimester MSc (AI & ML)

# Advanced Machine Learning

## Department of Computer Science

## AUDIO GENRE CLASSIFICATION

by

Himanshu Gulechha (2348520)
Manashwy Padhi (2348528)
Purusharth Malik (2348542)

January 2024

CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

# CERTIFICATE

*This is to certify that the report titled* ***Audio Genre Classification*** *is a bona fide record of work done by* ***Himanshu Gulechha (2348520), Manashwy Padhi (2348528), Purusharth Malik (2348542)*** *of CHRIST(Deemed to be University), Bangalore, in partial fulfillment of the requirements of II Trimester of Msc Artificial Intelligence and Machine Learning during the year 2023-24.*

**Course Teacher**

Valued-by: (Evaluator Name & Signature)
1.

2.

Date of Exam:

# Table of Contents

**Page Number**

**Team Details**

| **Reg. no** | **Name** | **Summary of tasks performed** |
|---|---|---|
| **2348520** | Himanshu Gulechha | Desicision Tree Classifier, Random Forest Classifier, KNN Classifier, SVM Classifier |
| **2348528** | Manashwy Padhi | EDA, Preprocessing, LDA, PCA, Clustering, Bagging, Boosting |
| **2348542** | Purusharth Malik | Dense and Convolutional Models, Evaluation and Comparison of all the models |

## 1. Abstract

This research proposal aims to delve into the domain of audio genre classification, employing the widely recognized GTZAN dataset. The study focuses on leveraging diverse machine learning models, both supervised and unsupervised, to enhance the accuracy of genre classification in audio content.

The primary objective of the research is to investigate the effectiveness of various machine learning techniques in the audio genre classification task. The proposed methodology employs traditional supervised models such as Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (KNN). Additionally, the study explores unsupervised models like clustering algorithms to uncover patterns within the audio data.

Preliminary experiments have revealed promising results, with particular emphasis on the outstanding performance of KNN and a deep learning approach utilizing dense layers exclusively. The KNN model demonstrated remarkable accuracy in capturing nuanced distinctions between different music genres, while the deep learning architecture achieved an impressive accuracy of up to 92%.

To validate and extend these findings, the research will systematically evaluate and compare the performance of these models on a larger scale, incorporating a comprehensive set of evaluation metrics. Furthermore, the study aims to explore ensemble methods and transfer learning techniques to enhance the robustness and generalization capabilities of the models.

This research has the potential to contribute significantly to the field of audio genre classification by identifying the most effective models for this specific task. The outcomes can be applied in various domains, including music recommendation systems, content categorization, and audio content organization, ultimately improving user experience and system efficiency.

## 2. Introduction

Audio Genre Classification (AGC) has been a longstanding challenge in music and audio signal processing, playing a pivotal role in content organization, recommendation systems, and the overall user experience in digital media. Traditionally, human experts and rule-based systems were employed for genre categorization, relying on manual annotation and subjective judgment. However, with the advent of machine learning (ML) techniques, particularly in recent years, the landscape of audio genre classification has witnessed a transformative evolution.

The endeavor to classify audio content by genre dates back to the early days of digital music libraries when the need for automated categorization became evident with the exponential growth of digital audio files. The GTZAN dataset, a widely recognized benchmark in AGC, has played a crucial role in shaping research progress in this domain. It comprises 1,000 audio clips, each 30 seconds long, spanning ten distinct genres. Researchers have extensively utilized this dataset to develop and benchmark various algorithms due to its diversity and comprehensiveness.

Integrating machine learning methodologies into AGC has revolutionized the field, providing automated solutions surpassing rule-based systems' limitations. Supervised learning approaches, in particular, have gained prominence by enabling models to learn complex patterns and features directly from the data. Techniques such as Support Vector Machines, Decision Trees, Random Forests, and K-Nearest Neighbors have been widely explored in the pursuit of accurate genre classification.

However, recent advancements in deep learning have introduced a paradigm shift in AGC. The application of neural networks, especially deep learning architectures, has demonstrated unparalleled capabilities in capturing intricate patterns and hierarchical representations inherent in audio data. Exploring deep learning techniques, such as dense and convolutional layers, has shown promising results, achieving remarkable accuracy in genre classification tasks.

This research proposal seeks to build upon the historical progression of AGC and the transformative impact of machine learning techniques, specifically focusing on utilizing the

GTZAN dataset. By employing a diverse range of machine learning models, including supervised and unsupervised approaches, this study aims to enhance the accuracy and robustness of AGC systems, contributing to the ongoing evolution of this dynamic field.

## 3. Data Pre-processing and Exploration

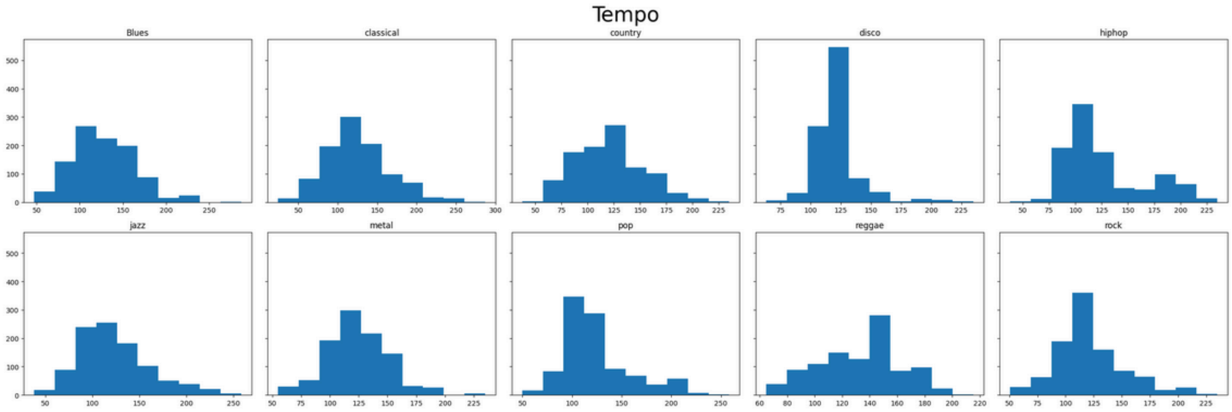
### 3.1 Data understanding and exploration

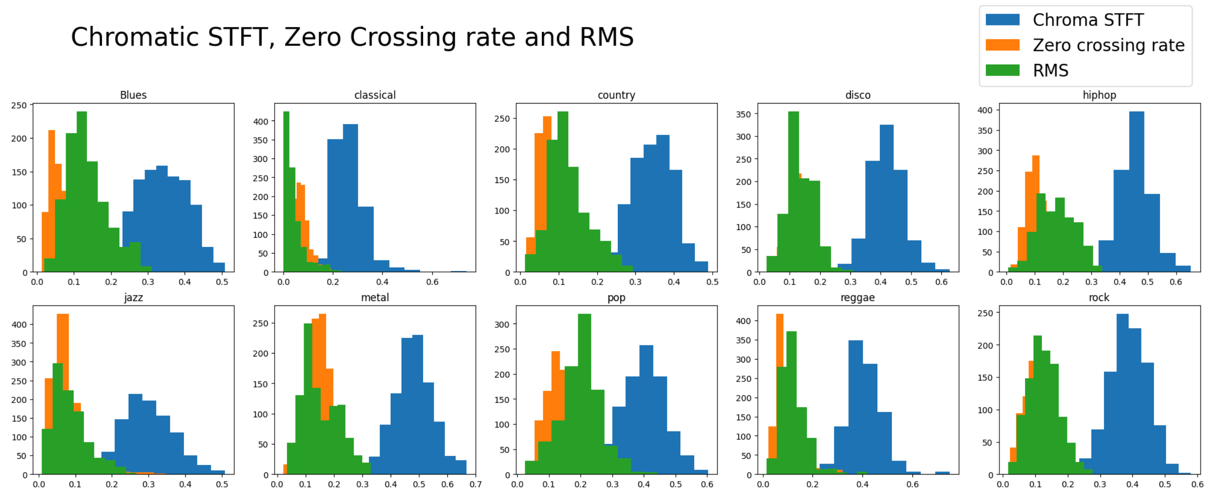
The dataset consists of 1,000 audio clips, covering a spectrum of 10 distinct genres. The songs were split before into 3-second audio files (this way, increasing 10 times the amount of data we fuel into our classification models). The genres include blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock, making GTZAN an ideal benchmark for evaluating the efficacy of machine learning models in genre classification.

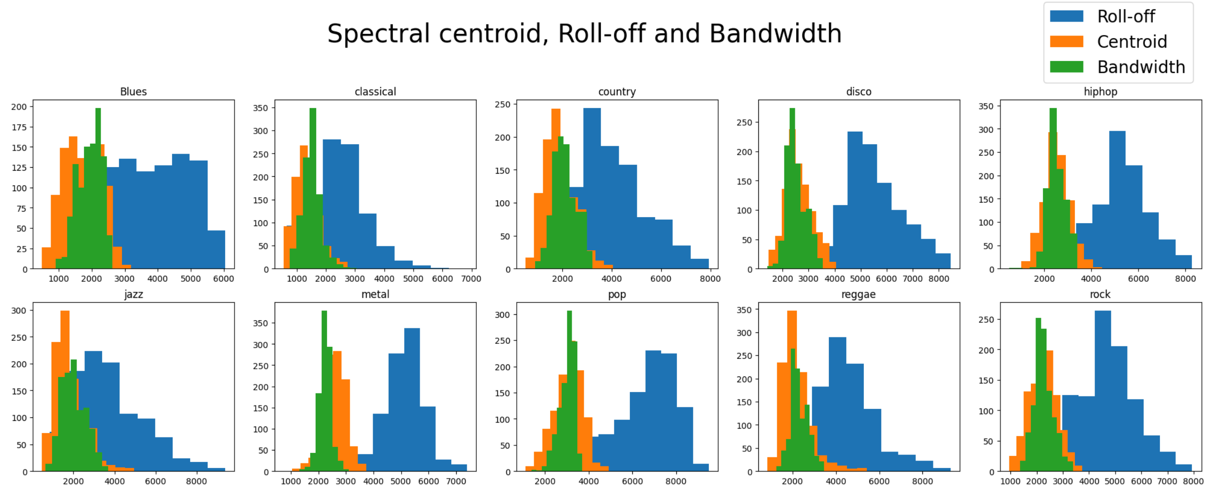
Features available in the dataset:

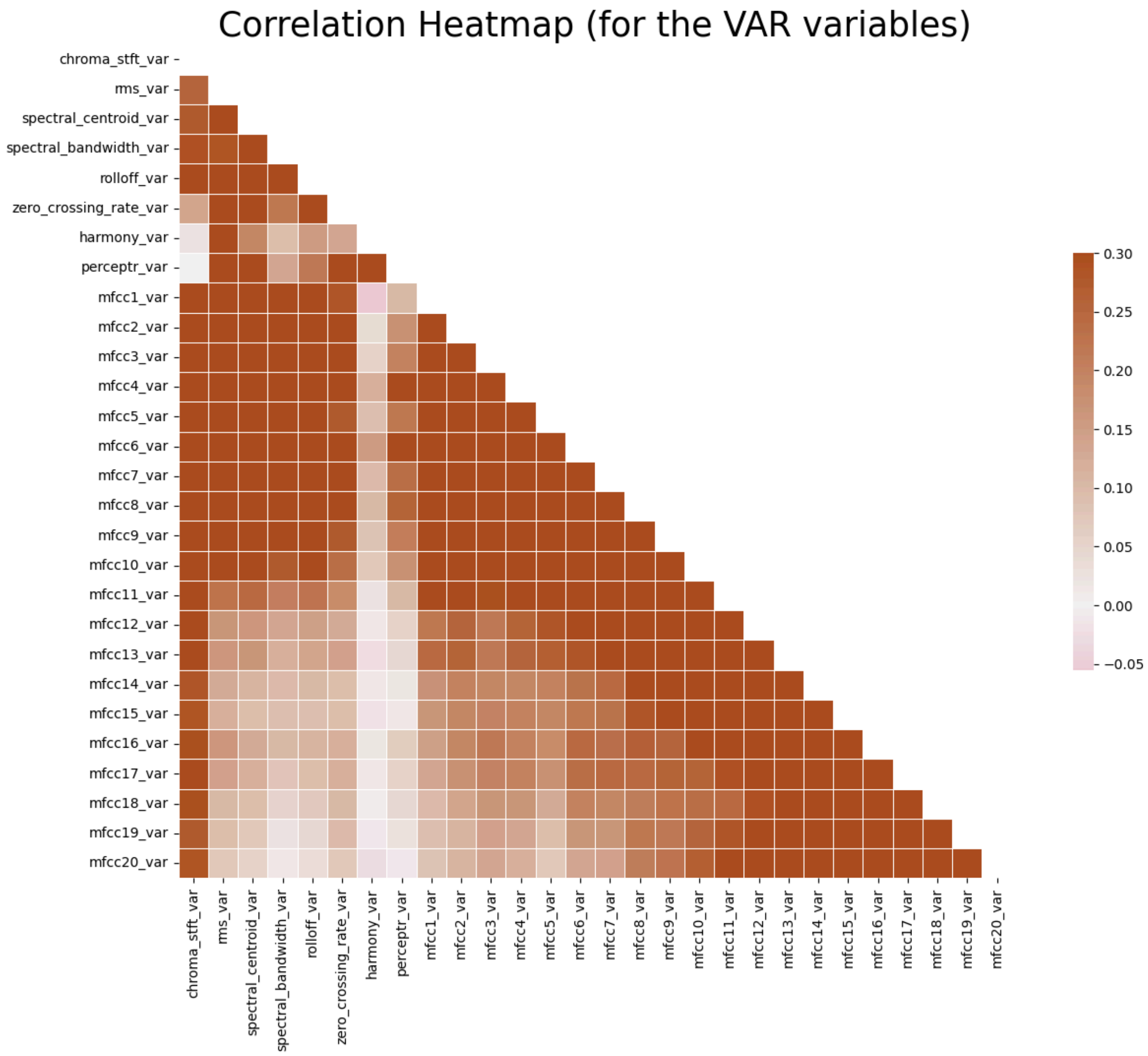
1. Chroma - Mean and Variance
2. RMS - Mean and Variance
3. Spectral Centroid - Mean and Variance
4. Spectral Bandwidth - Mean and Variance
5. Rolloff - Mean and Variance
6. Zero Crossing Rate - Mean and Variance
7. Harmony and Perceptual - Mean and Variance
8. Tempo
9. MFCCs - Mean and Variance

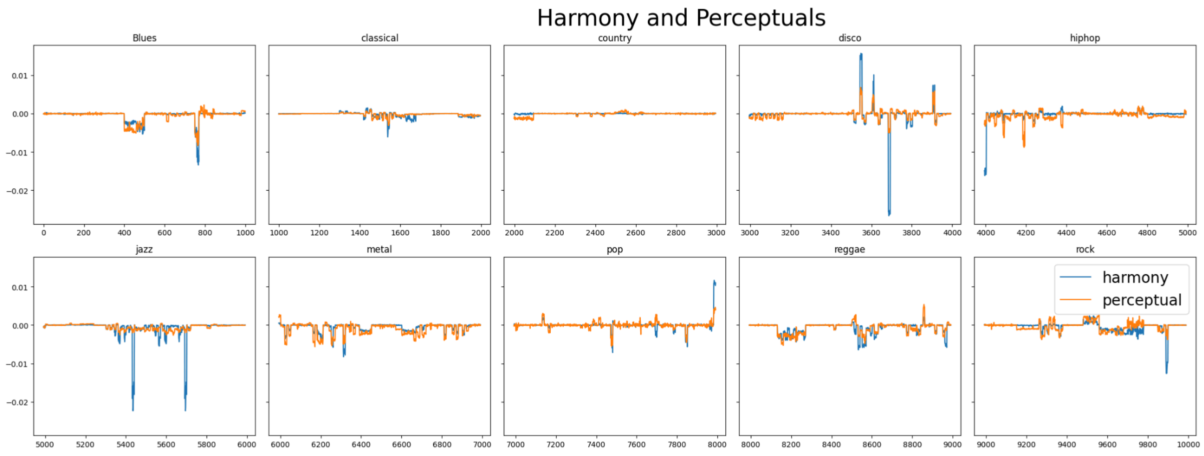
Then, we do the following data exploration:




Chromatic STFT, Zero Crossing rate and RMS
Chroma STFT
Zero crossing rate
RMS
Blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock


Spectral centroid, Roll-off and Bandwidth
Roll-off
Centroid
Bandwidth
Blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock


Correlation Heatmap (for the VAR variables)
chroma_stft_var
rms_var
spectral_centroid_var
spectral_bandwidth_var
rolloff_var
zero_crossing_rate_var
harmony_var
perceptr_var
mfcc1_var
mfcc2_var
mfcc3_var
mfcc4_var
mfcc5_var
mfcc6_var
mfcc7_var
mfcc8_var
mfcc9_var
mfcc10_var
mfcc11_var
mfcc12_var
mfcc13_var
mfcc14_var
mfcc15_var
mfcc16_var
mfcc17_var
mfcc18_var
mfcc19_var
mfcc20_var
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05


Harmony and Perceptuals
Blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
harmony
perceptual
0.01
0.00
−0.01
−0.02
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
4800
5000
5200
5400
5600
5800
6000
6200
6400
6600
6800
7000
7200
7400
7600
7800
8000
8200
8400
8600
8800
9000
9200
9400
9600
9800
10000

### 3.2 Data cleaning and handling missing values

Since there are no missing values, no data cleaning is required.

### 3.3 Data integration and feature engineering

Since we have a CSV file of the audio files and images of the corresponding spectrograms, there is no need of data integration and feature engineering.

## 4. Algorithm Implementation

### 4.1 Algorithms Implemented

### 4.1.1 Supervised Algorithms

- **Random Forest Classification**

- Ensemble Learning: Random Forest is an ensemble learning method that combines multiple decision trees to make more accurate and robust predictions. By aggregating the outputs of individual trees, Random Forest mitigates overfitting and improves generalization performance, contributing to enhanced accuracy in Audio Genre Classification (AGC).

- Feature Importance: Random Forest measures feature importance based on how much each feature contributes to the model's decision-making process. This is valuable in AGC, helping to identify the most relevant audio features for genre classification. By considering feature importance, the model can focus on the key attributes that distinguish between different genres.

- Robustness and Generalization: Random Forest is known for its robustness to noise and outliers in the data. The aggregation of predictions from multiple trees helps to smooth out individual errors and increases the overall reliability of the model. This robustness makes Random Forest particularly suitable for AGC tasks where audio datasets may contain variations, distortions, or irregularities.

After Hyperparameter Tuning, we get the following model:

criterion - entropy
max_depth - 8
max_features - sqrt
n_estimators - 1000

### - Support Vector Classification

- Effective in High-Dimensional Spaces: Support Vector Classifiers (SVC) are particularly effective in high-dimensional spaces, making them well-suited for Audio Genre Classification (AGC) tasks where audio features may span a large feature space. The algorithm is adept at identifying complex decision boundaries in such high-dimensional settings, contributing to accurate genre classification.

- Robust to Overfitting: SVCs incorporate the concept of a margin, which represents the distance between the decision boundary and the nearest data point of any class. By maximizing this margin, SVCs aim to find a hyperplane that generalizes well to unseen data, reducing the risk of overfitting. This robustness is beneficial in AGC, especially when dealing with limited labeled audio data.

- Kernel Trick for Non-Linear Decision Boundaries: SVCs can leverage the kernel trick to map the input features into a higher-dimensional space, allowing them to handle non-linear decision boundaries. This capability is crucial in AGC, where the relationships between audio features and genres may be complex and non-linear. The choice of an appropriate kernel (e.g., radial basis function kernel) enables SVCs to capture intricate patterns in the data, enhancing their classification performance.

After Hyperparameter Tuning, we get the following model:

C - 100

gamma - 0.01

kernel - rbf

### - K-Nearest Neighbours

- Localized Decision Making: K-Nearest Neighbors (KNN) is a non-parametric, instance-based algorithm that makes decisions based on the majority class of the k-nearest data points. In the Audio Genre Classification (AGC) context, KNN considers an audio sample's similarity to its nearest neighbors, allowing it to capture localized patterns and adapt to the specific characteristics of different genres.

- Simple and Intuitive: KNN is a straightforward algorithm that is easy to understand and implement. Its simplicity makes it particularly suitable for AGC tasks, where a clear understanding of the decision-making process is valuable. KNN's intuitive concept of classifying samples based on the classes of their closest neighbors adds to its appeal, especially when interpretability is a priority.

- Adaptability to Feature Spaces: KNN is versatile in handling different types of feature spaces, whether continuous, discrete, or a mix of both. In AGC, where audio features may have diverse characteristics, KNN's adaptability to various feature types makes it a flexible choice. It can effectively capture the relationships between audio samples in a high-dimensional feature space.

After Hyperparameter Tuning, we get the following model:

metric - manhattan
n_neighbours - 3

### 4.1.2 Unsupervised Algorithms

- **KMeans**

- Unsupervised Clustering: KMeans is an unsupervised machine learning algorithm used for clustering data into distinct groups based on similarity. In the context of Audio Genre Classification (AGC), KMeans can be applied to identify inherent patterns and group similar audio samples together without the need for labeled genre information. This can be useful for exploratory analysis and uncovering natural clusters within the data.

- Simplicity and Efficiency: KMeans is a simple and computationally efficient clustering algorithm. It iteratively assigns data points to clusters based on the mean of their features, making it computationally less demanding compared to more complex algorithms. Its efficiency is advantageous in scenarios where quick insights into audio data structures are needed, making it a valuable tool in the initial stages of AGC research.

- Feature Representation Exploration: KMeans can be employed to explore different feature representations of audio data. By clustering audio samples based on their features, researchers can gain insights into how various features contribute to forming distinct groups. This exploration aids in understanding the diversity within the dataset and can guide the selection of relevant features for subsequent supervised learning models in AGC.

- **BIRCH**

- Incremental and Scalable: BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an incremental and scalable clustering algorithm. It processes data points one at a time, making it suitable for streaming or large datasets. In Audio Genre Classification (AGC) context, BIRCH can efficiently handle continuous streams of audio data or large audio databases.

- Hierarchical Clustering: BIRCH employs a hierarchical clustering approach, organizing data points into a tree-like structure called the Clustering Feature Tree (CF Tree). This hierarchical representation can be beneficial in AGC by providing insights into the relationships between different clusters at multiple levels, allowing for a more nuanced understanding of the diversity within audio genres.

- Adaptability to Varying Densities: BIRCH adapts well to clusters with varying densities, making it robust in scenarios where audio samples within different genres may exhibit different levels of compactness. This adaptability allows BIRCH to effectively capture both dense and sparse regions in the feature space, which can be advantageous in AGC, where audio samples may have diverse characteristics.

### 4.1.3 Dimensionality Reduction Algorithms

- **Principal Component Analysis (PCA)**

- Dimensionality Reduction:
  - Principal Component Analysis (PCA) is a technique for reducing the dimensionality of data while retaining most of its variability. In the Audio Genre Classification (AGC) context, PCA can be applied to reduce the number of features in the audio data, simplifying the computational complexity of subsequent models and potentially highlighting the most relevant features for genre classification.
- Decorrelation of Features:
  - PCA transforms the original features into a new set of uncorrelated features called principal components. This decorrelation can be beneficial in AGC, where certain audio features may be correlated or redundant. By capturing the directions of maximum variance in the data, PCA provides a more compact and uncorrelated representation, potentially improving the efficiency of subsequent machine learning models.
- Visualization of Data:
  - PCA can be used for visualizing high-dimensional data in a lower-dimensional space. In AGC, this visualization can help researchers and practitioners gain insights into the distribution of audio samples across different genres. By representing the data in a reduced-dimensional space, PCA facilitates

exploring and understanding the inherent structure and relationships within the audio dataset.

- **Linear Discriminant Analysis (LDA)**

- Supervised Dimensionality Reduction: Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction technique that aims to maximize the separation between classes while preserving information within each class. In the context of Audio Genre Classification (AGC), LDA can be applied to reduce the dimensionality of the feature space in a way that maximizes the differences between genres, potentially improving the performance of subsequent classification models.

- Classification-Focused Projection: Unlike PCA, which focuses on maximizing variance, LDA considers explicitly the differences between classes. LDA seeks to create a feature space where different genres are well-separated by finding a projection that maximizes the ratio of between-class and within-class variance. This is particularly beneficial in AGC, where the goal is to distinguish between various audio genres.

- Handling Multi-class Classification: LDA is naturally suited for multi-class classification problems, making it applicable to AGC tasks with multiple genres. It provides a discriminative feature space that optimally separates two classes and multiple classes simultaneously. This makes LDA a relevant choice for modeling the relationships between different audio genres in a unified manner.

#### 4.1.4 Deep Learning Models

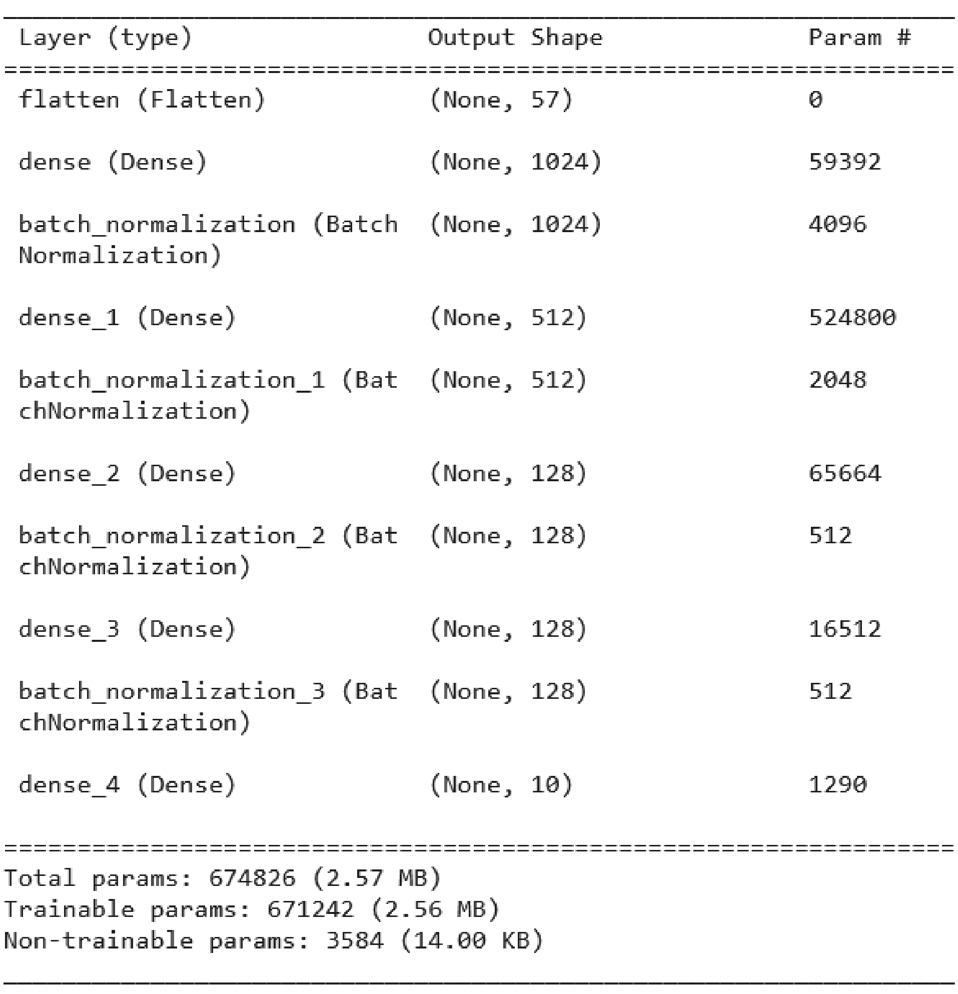
- **Fully Connected Model**

- End-to-End Representation Learning: Fully connected deep learning models, often implemented as dense neural networks, are capable of learning complex hierarchical representations directly from the raw audio data. In the context of Audio Genre Classification (AGC), these models can automatically extract hierarchical features that capture both low-level and high-level characteristics of audio samples, potentially improving the model's ability to discern genre-specific patterns.

- Non-linear Mapping: Fully connected deep learning models introduce non-linear activation functions, allowing them to capture intricate relationships and patterns within the audio data. This non-linear mapping is crucial in AGC, where the relationships between audio features and genres are often complex and non-linear. The model can learn nuanced representations that may be challenging for linear models to capture effectively.

- Adaptability to Complex Data: Deep learning models, with fully connected layers, are highly adaptable to the complexity and variability present in audio data. These models can automatically learn and adapt to diverse features and patterns within different genres. The flexibility of deep learning architectures makes them well-suited for AGC tasks, where the characteristics of audio samples can vary widely across genres and over time.

  Architecture:



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 flatten (Flatten)           (None, 57)                0

 dense (Dense)               (None, 1024)              59392

 batch_normalization (Batch  (None, 1024)              4096
 Normalization)

 dense_1 (Dense)             (None, 512)               524800

 batch_normalization_1 (Bat  (None, 512)               2048
 chNormalization)

 dense_2 (Dense)             (None, 128)               65664

 batch_normalization_2 (Bat  (None, 128)               512
 chNormalization)

 dense_3 (Dense)             (None, 128)               16512

 batch_normalization_3 (Bat  (None, 128)               512
 chNormalization)

 dense_4 (Dense)             (None, 10)                1290

=================================================================
Total params: 674826 (2.57 MB)
Trainable params: 671242 (2.56 MB)
Non-trainable params: 3584 (14.00 KB)
_________________________________________________________________
```

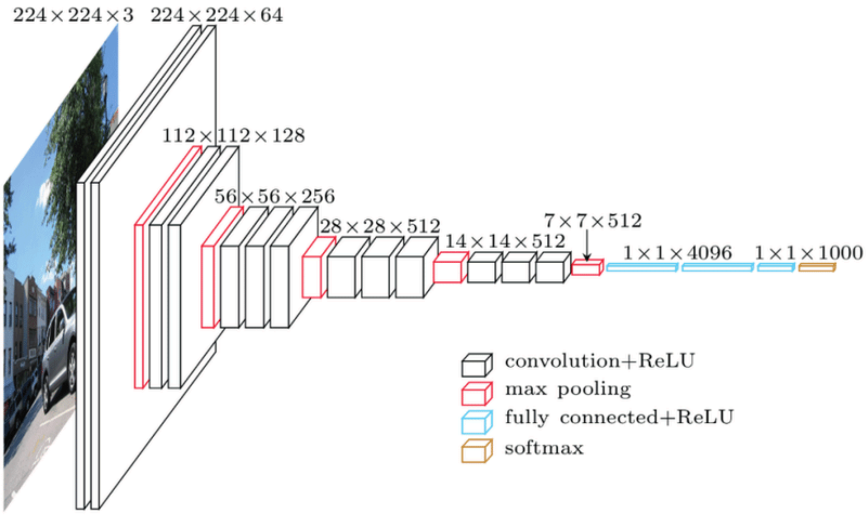
- **Transfer Learning Using VGG16**

- Feature Extraction and Transferability: VGG16, a deep convolutional neural network architecture, has been pre-trained on large-scale image classification tasks. In transfer learning for Audio Genre Classification (AGC), the early layers of VGG16 can be repurposed as feature extractors for audio spectrograms. The learned visual features, though initially intended for images, often demonstrate transferability to other domains, including audio. This allows leveraging the knowledge gained from image classification tasks to enhance the representation of audio features.

- Reduced Training Time and Data Requirements: Transfer learning with VGG16 can significantly reduce the training time and data requirements for an AGC model. By utilizing the pre-trained weights of VGG16 as a starting point, the model starts with features that are already capable of capturing generic patterns. This is especially advantageous when dealing with limited labeled audio data, common in AGC scenarios. The pre-trained model provides a head start in learning relevant representations, requiring less data and computation for fine-tuning on the specific AGC task.

- Improved Generalization and Robustness: Transfer learning using VGG16 enhances the generalization and robustness of AGC models. The knowledge captured by VGG16 from diverse image datasets can help the model generalize well to different audio genres. The hierarchical and abstract features learned by VGG16's deep layers contribute to a more robust representation, enabling the AGC model to better capture complex patterns and variations within audio data. This leads to improved performance when compared to training a model from scratch, particularly in scenarios with limited labeled audio data.

Architecture of VGG16:



### 4.2 Correct parameter tuning

Hyperparameter Tuning was done for all the ML models using appropriate parameter values. For deep learning models, Keras Tuner was used to fine-tune the number of neurons in the hidden layers.

### 4.3 Efficient coding and algorithm execution

The deep learning models were executed on a single P-100 GPU for faster execution and greater accuracy.

## 5. Model Evaluation and Performance Analysis

### 5.1 Evaluation metrics and performance assessment

Following are the evaluation metrics that were followed to compare different models:

1. Accuracy:
   - Definition: The proportion of correctly classified instances among the total instances.
   - Use Case: Accuracy provides an overall measure of model performance. However, it may not be suitable in cases of imbalanced datasets, where one genre is more prevalent than others.
2. Precision, Recall, and F1-Score:
   - Precision:
     - Definition: The ratio of correctly predicted positive observations to the total predicted positives.
     - Use Case: Precision is valuable when the cost of false positives is high, such as misclassifying a non-relevant audio sample as belonging to a specific genre.
   - Recall (Sensitivity or True Positive Rate):
     - Definition: The ratio of correctly predicted positive observations to the total actual positives.
     - Use Case: Recall is important when the cost of false negatives is high, such as missing a relevant audio sample from a specific genre.
   - F1-Score:

        - Definition: The harmonic mean of precision and recall, providing a balance between the two metrics.
        - Use Case: F1-score is useful when there is a need to strike a balance between precision and recall.

3. Confusion Matrix:
    - Definition: A table showing the true positive, true negative, false positive, and false negative values.
    - Use Case: The confusion matrix provides a detailed breakdown of model performance, helping to identify specific areas of improvement and error types.

### 5.2 Comparative Analysis of Different Models

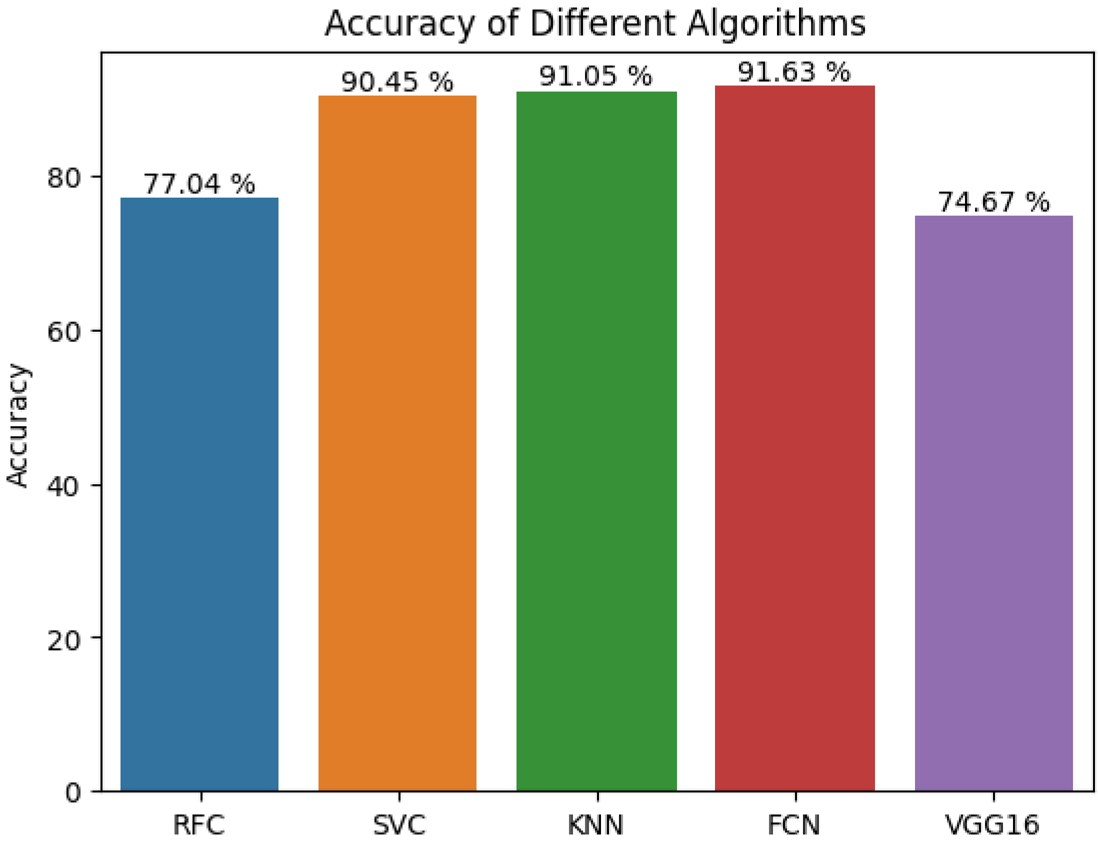
Since the classes are balanced and sampling has been done in a stratified manner, accuracy gives an accurate representation of how well a model performs on the GTZAN dataset.

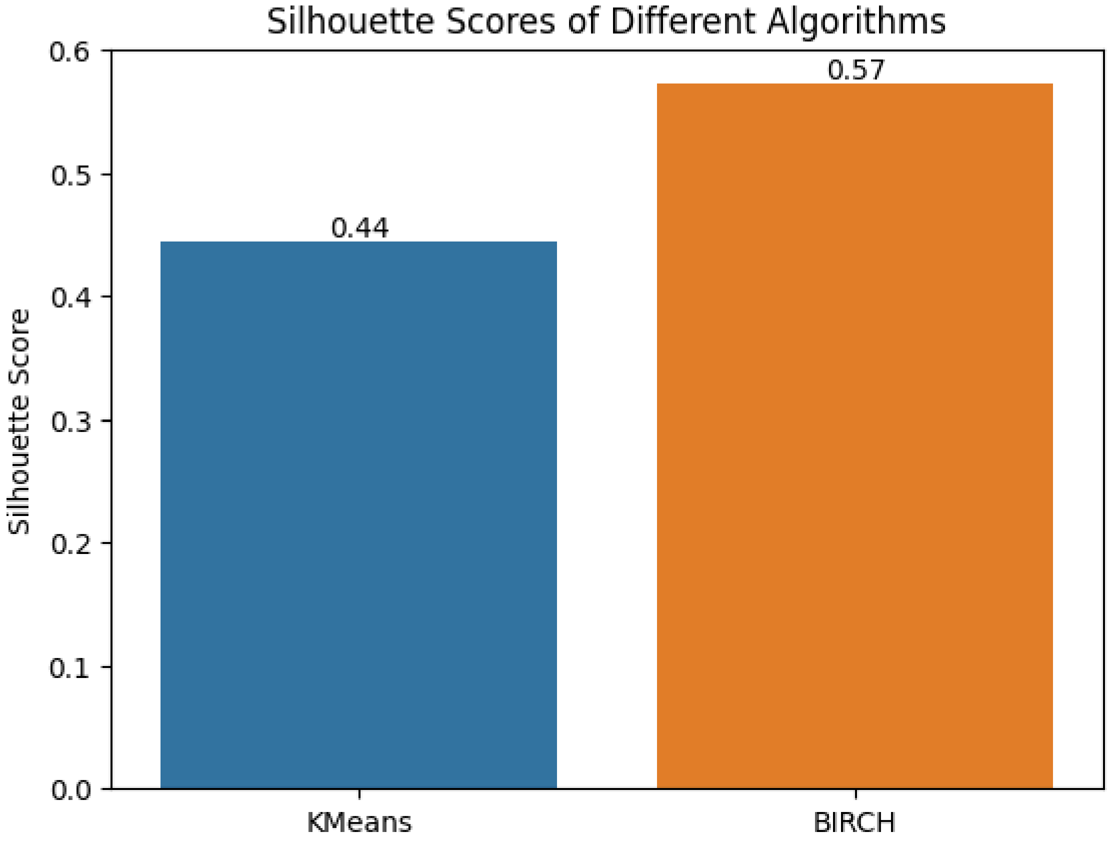
Note - For unsupervised algorithms, silhouette score will be calculated.

| **S. No.** | **Algorithm** | **Accuracy** |
|---|---|---|
| **1.** | Random Forest Classifier | **77.04%** |
| **2.** | Support Vector Classifier | **90.45%** |
| **3.** | K-Nearest Neighbours | **91.05%** |
| **8.** | Fully-Connected Network | **91.63%** |
| **9.** | Transfer Learning Using VGG16 | **74.67%** |

Unsupervised Algorithms:

| S. No. | Algorithm | Silhouette Score |
|---|---|---|
| **1.** | K-Means with PCA | **0.4440** |
| **2.** | BIRCH | **0.5724** |





### 5.3 Insightful interpretation of results

- Supervised Learning

- K-Nearest Neighbors (KNN) and Fully-Connected Network achieved the highest accuracy, both exceeding 91%. This suggests that these algorithms are well-suited for the task of AGC and are able to effectively learn the patterns in the data.
- Support Vector Classifier (SVC) performed significantly better than Random Forest Classifier, suggesting a more suitable model for this specific task. This could be due to the fact that SVC is better at handling non-linear data, while Random Forest Classifier is more suited for linear data.
- Transfer Learning using VGG16 underperformed compared to other algorithms, indicating that the pre-trained model might not be well-suited for the AGC task. This

is likely because the VGG16 model was trained on a different dataset (ImageNet) that is not closely related to the AGC task.

- Unsupervised Learning

- BIRCH achieved a higher Silhouette Score compared to K-Means with PCA, suggesting better cluster separation. This means that the clusters produced by BIRCH are more distinct and well-defined than the clusters produced by K-Means with PCA.
- However, the interpretation of Silhouette Scores depends on the specific dataset and task. A higher score might not always be directly translatable to better clustering quality. In some cases, a lower Silhouette Score might actually be better, if it indicates that the clusters are more compact.

**Overall, the results suggest that KNN, Fully-Connected Network, and SVC are the best performing algorithms for the task of AGC. BIRCH also performed well in the unsupervised learning task.**

**Note - It is important to remember that VGG16 was trained on image data i.e., it can not be directly compared all the other algorithms as the dataset was different.**

## 6. References

1. Dataset - https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification
2. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv*. /abs/1409.1556
3. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2019). A Comprehensive Survey on Transfer Learning. *ArXiv*. /abs/1911.02685
4. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
5. Geÿron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly.
6. Meinard Müller (2021). Fundamentals of Music Processing Using Python and Jupyter Notebooks. Springer Cham.