Detecting Music Similarity - A Novel Framework Combining Metadata and Content-Based Similarity Measurements Using Deep Learning

Lavonnia Newman[1], Dhyan Shah[1], Chandler Vaughn[1], Faizan Javed[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Adjunct Lecturer, Data Science, Southern Methodist University,
Dallas, TX 75275 USA
{lavonnian, dhyans, cvaughn, fjaved}@smu.edu

Abstract. Music is incorporated into our daily lives whether intentional or unintentional. It evokes responses and behavior so much so there is an entire study dedicated to the psychology of music. Music creates the mood for dancing, exercising, creative thought or even relaxation. It is a powerful tool that can be used in various venues and through advertisements to influence and guide human reactions. Population segmentation can be achieved through the musical genres. Identification with genres such as classical, pop, rock, country, gospel can be used as avenues to connect with the targeted audience to influence purchases, associate with candidates running for office, or even trigger long ago memories. Classification of music into one genre or across genres increases the targeted audience pool, hence detecting similarities and classifying music into those various groups is a useful tool for companies and organizations. With the age of the Internet, it is now easier than ever to access, distribute, modify, and create derivative musical works. Yet programmatically assigning genre classification, and determining music similarity is still a nacent field of study. Through the course of this paper, we explore various approaches to detecting music similarity between songs. We then propose a unique, data-driven, approach to detecting music similarity across a diverse feature space.

1 Introduction

Almost everyone you speak with loves Music. It evokes emotion and enriches our lives. In many ways it is a core component to our human existence, and it is embedded in our daily lives so much that it can often go unnoticed. Unintentional music exposure occurs through commercial advertisements, elevators, bars, restaurants, stores, trains, planes, movies, sporting events and many more venues too numerous to list. It is used to set the mood for the occasion such as dancing, aerobic exercise, praise and worship, or even relaxation or work [1]. When combining intentional music listening with background music, the average American is exposed to music more than 30% of their waking hours per day [1].

Individual musical tastes vary considerably. Musical genres such as rock, classical, pop, country, folk, easy listening, and gospel have been used to segment society, to varying degrees of success. But, over the past few decades, we have seen an explosion of musical categories. Spotify, for example, taunts their classification of 1,387 sub-genres [21]. A sub-genre is genre within a genre that can be defined by geographical region, musical techniques, or even cultural context. This categorization explosion has largely been driven by massive technological shifts in computing, networking, data processing, and music production equipment. The new categories like neo mellow, grave wave, metropolis, bhangra and nu gaze may not trigger a response from the average listener if you question them on their familiarity with songs in the genre. Anyone with a few thousand dollars can produce an album in their garage and distribute it via the Internet. This, coupled with considerable business model changes for the music industry towards *streaming*, have created a rapidly advancing need for automation and intelligence in categorization and analysis of musical works that exceed the old standards of musical genre classification. The landscape of what is considered music has changed and detecting the similarities in music is an extraordinary feat. One of the chief data scientists at Spotify created a visual display of the subtle changes or sonic differences in sound on the website *everynoise.com*. On this website you can see music sounds mapped along the spectrum of mechanical to organic to atmospheric. This website is a fantastic depiction of the complexity of musical classification due to the vast contrast in sounds.

Motivation: With the age of digital technology, it is now easier than ever to explore inherent musical genre and its characteristics. Musical genres are categorical labels created by artists, listeners, and producers to characterize music. These labels are carefully crafted descriptions of audio instrumentation, rhythmic structure, and harmonic content which forms the overall musical genre category. Music services today attempt to streamline music choices to targeted audiences, largely to the benefit of the service provider. There are opportunities to identify *music crossover* where a musical work can be classified across two or more genres, thus providing a potentially larger target audience for an artist or genre. Genre annotation is critical to this process and is unfortunately largely performed manually in the industry. Given the state of computing and data science, we believe that we can build on the current state of the art music similarity research, as well as existing machine learning models to improve music similarity search. [17]

Commented [VC1]: Remove before final draft print

Machine learning classification of musical genre can provide a semantic meta-data framework on feature extraction for a content-based analysis of any music. Most music forensics analysis has its roots in signal processing, but many of the practices are still nascent and evolving. Have you ever heard of a “forensic musicologist”? The term describes someone that analyzes music not as a singer, composer or writer, but as someone that is focused on analyzing and detecting structural similarity in music. Our research seeks to describe a framework of feature analysis via machine learning and apply it to music genre classification and music similarity. [18]

The breadth of research yet devoted to music similarity measurement is relatively finite. This is especially true when considering the proliferation and acceleration of adoption of digital music, and the increasing practice of sampling during the creative process for new works. Zhang, et al. classified research areas in music similarity as generally belonging to the categories of “Metadata-based similarity, content-based similarity, or semantic description-based similarity” [7]. Given the nascent field, we chose to restrict our research to content-based methods. This rich area provides a rich feature playground, with low-level, but highly dimensional, features to work from.

Tangential research exists that we can learn from as well. This includes studies on how chroma features and fingerprinting can be used to identify cover songs, arguably a related problem to sampling. However, few approaches specifically target building algorithmic approaches to detecting music similarity, or sampling as “a song within [another] song” [7, 8].

Problem Statement: How to evaluate, model, and analyze music similarity remains one of the more challenging and diverse research areas in audio analysis today. Through the course of this research, we consider various approaches to detecting similarity in target songs, and then utilize the rich feature space of raw

Commented [VC2]: Remove before final draft print audio, along with machine learning, to propose a unique, data-driven, approach to detecting music similarity.

The remainder of this paper is organized as follows:

- Section 2 provides an overview of how music is currently analyzed, and the associated properties that are measured,
- Section 3 covers ethics and ethical implications for this work,
- Section 4 provides a survey of techniques that are important to understand
- Section 5 covers our proposed framework and approach for music plagiarism detection,
- Section 6 describes the empirical results to our testing and experimentation,
- Section 7 suggests future avenues for study,
- Section 8 outlines lessons learned from this research, and;
- Section 9 concludes our findings.

2 Music Analysis & Properties

To understand how we intend to analyze audio for similarity, its important for us to take a moment to educate the reader to the feature-rich space that audio provides. While the terms *unstructured data* typically elicts ideas of numerous data points or text attributes, on several topics, with no obvious way to statistically analyze them, in this case we use it to definitively describe raw audio. Both the process of feature extraction, and the features themselves, deserve some review. While we will not attempt to cover the full sphere of audio analysis, we provide the below tutorial on audio analysis as a baseline starting point for the reader to further understand our work, and to appreciate the complexities for audio analysis.

2.1 Tutorial – Background Definitions

Most readers will relate to common audio concepts such as frequency and amplitude. *Frequency* represents the speed of a vibration, which thus determines *pitch*. *Amplitude*, in contrast, is the size of that vibration. So if frequency determines pitch, amplitude can be thought of deterministic to how loud that pitch is. The spectrum of human-detectable hearing ranges to 12 distinct pitches, through various octaves. This, at its core, makes up the basis of a sort of DNA for music and coupled with temporal, rhythmic and melodic features, it is the basis of modern western music that we know today. This also provides a basic framework for understanding some of the audio feature extraction techniques we describe below.

One principle area of audio feature extraction centers on a term described as *chroma*. Chroma is representative of tonal and pitch content. In literal terms, it is the "color" of a musical pitch, decomposed such that it is octave-invariant into 12 pitch classes. Chroma features help capture musical characteristics in a condensed, potentially visual, form while "being robust to changes in timbre and instrumentation" [14]. Chroma features are typically extracted from raw audio content utilizing Short Time Fourier Transforms, Constant Q Transforms, and normalized Chroma Energy methods, as shown in **Fig. 1. Chroma Feature Extraction Process.**

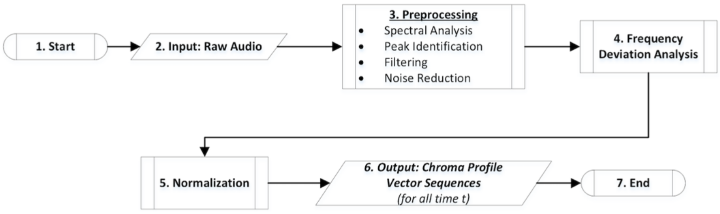

Fig. 1. Chroma Feature Extraction Process

It is logical to think that for any time *t* in a local time window of a song, it would have distinct chroma features. If we think of building these chroma feature windows across all windows for the song, we can build a representative pitch profile of the song over all 12 pitch chroma bands. When combined, this creates what is known as a *chromagram* for the song (see Fig. 2. Chromagram Example). Chromagrams are sometimes displayed as the squared magnitude of the Fourier coefficients at each section. In these cases, it is known as a *spectagram*. This spectagram is one way to densely represent a songs pitch structure.

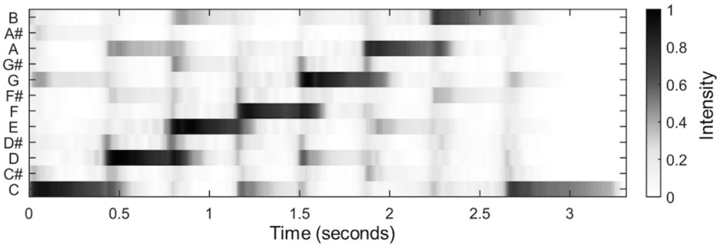

Fig. 2. Chromagram Example

Chromagrams can be thought of as one of the many building block features to audio processing. We cover it here only to instruct the reader on the type of processing that is done in order to extract features from audio. Several different pre-processing and post-processing techniques can be applied to raw audio to yield varying spectral, rhythmic, temporal, and melodic features for raw audio. Some of these subsequent features are tabulated below. While an exhaustive review of how these features are calculated is beyond the scope of this paper, we encourage the reader to review references for these feature types to learn more. For those features we utilize directly, we will explain in line with their use.

Table 1. Example Acoustic Features

Category	Category Description	Feature Space
Timbral	Tonal texture	Harmonic Pitch-Class Profile Spectral Centroid Spectral Contrast Rolloff Low-Energy Mel-frequency Cepstral Coefficient
Temporal	Time domain signals	Zero Crossing Rate Autocorrelation Waveform Moments Amplitude Modulation (loudness)
Spectral	Musical characteristics by spectra	Auto-regressive features Spectral asymmetry Kurtosis Flatness Crest factors Slope Decrease Variation Frequency derivative of Constant-Q Octave-band signal intensities
Rhythmic	Musical timing	Beat histogram Rhythm strength Regularity Average tempo
Melodic	Melodic content	Pitch histogram

Some of the more common tools used today for music and speech recognition include mel-frequency cepstral coefficients (MFCC), and spectrogram analysis. Unfortunately, both of these methods are sensitive to additive noise and pitch dynamics. As a result, it is common to see log scaling as a post-processing step. (see Fig. 3. Spectogram Log Scale Example and Fig. 4. Mel-frequency Cepstral Coefficient (MFCC) Example)

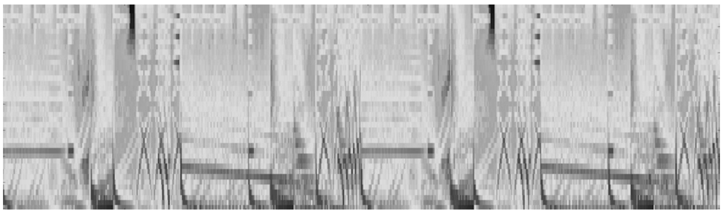

Fig. 3. Spectogram Log Scale Example

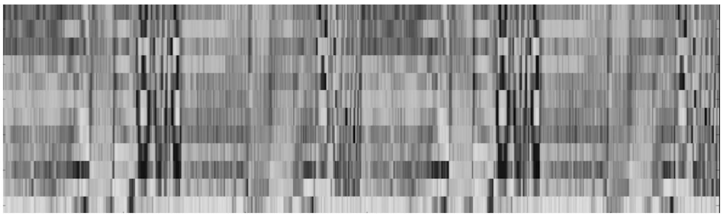

Fig. 4. Mel-frequency Cepstral Coefficient (MFCC) Example

2.2 Feature Extraction

Now that we have covered some of the definitions and concepts around the chromagram extraction process, let's walk through the process leading to feature extraction that will ultimately allow us to detect similarities between two musical compositions. As shown in Fig. 4. below the process begins with the music file, specifically a WAV file. While one might think the file format does not matter, it actually does. The various music file formats are used for particular purposes. The most popular file type MP3 is used for the distribution of music, downloads and posting on websites. It is compressed, uses little disk space, and offers a CD like quality that is appropriate for listening and multimedia presentations. A WAV file, however, is uncompressed and large in size. It too delivers CD quality sound, but it is better for music analysis and editing. WAV files can be looped together for repetition and can provide a seamless playback while MP3 files cannot. The first 10ms to 50ms of a MP3 snippet will always have a dead space that is due to the compression algorithm that created it. You cannot loop an MP3 sound bite without having those silent gaps. If comparing the concept of an audio WAV file to the video presentation. The WAV file would be considered high resolution while MP3 would be considered low resolution. Even though WAV file is the preferred format, related work does exist where MP3 files were used.

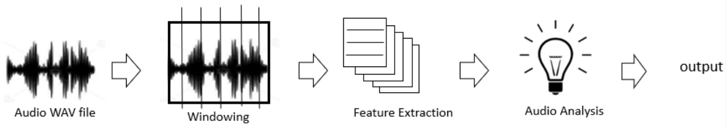

Fig. 4. Audio Analysis Process

We selected Essentia for analysis and feature extraction of the WAV file. Essentia is an open source library that was developed in C++ and wrapped in Python wrappers for the purpose of audio signal processing and audio music retrieval. Using the tools of Essentia we loaded the audio stream, which is the WAV file, and divided it into short audio segments. This process which is called windowing, results in a continuous sequence of blocks or chunks that will ultimately be subjected to audio signal analysis. The window function length is configurable and often is based on the length of a particular event of interest as would be the case for plagiarism.

After segmenting the music stream into audio chunks, the features are extracted. Features, which are the contextual data of an audio signal, are the foundation of research and development in audio signal processing. Features capture the physical and perceptual impact of the signal. The goal of feature extraction is to obtain the notable characteristics of a signal and converting them to coefficients. Once extracted these features can be used for data mining, data classification, statistical analysis and in our case similarity measures. In the following paragraphs we provide insight into the various descriptors and the associated features used for music information retrieval that form the analytical dataset for this project.

Timbral Descriptors - The features in this domain are related to the timbre of the music or the tone quality. Tonality in music is the organization of the note on a musical scale. It describes the sound's structure that is composed of harmonized related frequencies. The brightness feature would also be contained in this domain.

Temporal Descriptors - These features are unique in that they do not require the audio signal to be transformed. These computations occur on the original signal sample. Overall the temporal domain features capture the duration of the signals, the loudness, and the durations that cross a peak energy level. Common terms for capturing the loudness and energy levels are amplitude, which is the temporal structure of the signal, power which is also temporal but focused on the signal's power. One of the popular features in this category is the zero-crossing rate, which is the rate the signal goes from negative to zero to positive amplitude and vice versa, which is often using in classifying percussion sounds.

Spectral Descriptors – These features are also known as frequency-based features. This domain contains the largest group of audio features. The features are derived from an autoregression analysis or Short Term Fourier Transform (STFT) and describe the physical properties of the signal frequency. Autoregression features capture the results from a linear prediction analysis of the signal while the STFT captures a derivative from the signal's spectrogram.

Rhythm Descriptors - The rhythm features capture the organization of sonic events along a time axis.[REF] The beats per minute, the beat loudness, and other beat tracking features are found in this domain.

. Melodic Descriptors – The features in this category are related to the melody, pitch, chords, harmony, and tuning.

The Essentia package has a list of additional features that are noteworthy like danceability, fade detection, and dynamic complexity. It also contains a number of algorithms that allow for file manipulation, PCA and dynamic complexity. Essentia also comes with pretrained classifier models for various classifications like musical genre, ballroom music, moods, western, tonal, danceability, voice, gender, and timbre. Once the features are extracted using Essentia, A file containing the features and associated values is generated for each of the signal chunks. These files are now the data input files to be merged and exploratory data analysis to be performed on prior to engaging in any machine learning modeling.

3 Ethics

There are several ethical considerations in this analysis. Obviously in music there are the general concerns over plagiarism and copyright infringement. A use or close imitation of a substantial part of another musician's work claiming one's own music without a proper credit to original artist constitutes music plagiarism. Further, the music industry attempts to restrict classifications of "sound-a-likes," making it difficult for anyone to make a claim that their song sounds like another artist. These issues complicate technological evolution since finding music similarity is, by its definition, the process of finding songs that sound like other songs.

Further, from a researcher perspective, there are ethical considerations around bias. Researchers may tend to choose known or readily available algorithms and constructs to leverage in their research, which may inherently bias music similarity to the potential end detriment to certain artists music. Plus, music is still very much a human experience. Emotion, mood, and feeling are continuums and not necessary ordinal or categorical measurements. This has direct implications for ethical considerations [14].

3. Music Industry Ethical Considerations

Music similarity is only considered plagiarism if the original work has copyright protection. However, plagiarism does not necessarily result in a copyright infringement. Worldwide, the music industry is suffering from what might be considered as plagiarism, as end-users have instantaneous, easy access to digital media and contents over the internet, and can easily sample, re-sample, and modify existing artistic works by incorporating them into new derivative works.

While there are no defined standards to evaluate music similarity or plagiarism, the intellectual property laws in the U.S. fall short of clearly and legally defining the issue. The most recent Digital Millennium Copyright Act [14] defines the statute of limitation to 70 years after the death of the creator of a musical work, which has blurred the lines between the original song and an allegedly infringed song.

Venturing into music forensics and applying machine learning techniques, one must be cautious in evaluating music similarity. Ethically speaking, there can be three possible categories for musical similarity [16]:

1. Inspiration,
2. Coincidence, and;
3. Plagiarism/Copying/Wrongful appropriations.

Inspiration is considered a legitimate element of similarity in any form of artistic production. For centuries musicians have genuinely inspired from other variants of music (e.g. folk music) as long as inspiration is limited to only fragments of artistic bytes and synthesized into a musician's own perspective to create original music.

Additionally, there are only 12 notes in Western music and therefore coincidence of musical similarity between two songs have a non-zero probability. Thankfully, there are various other dimensions to a song. In legal terms, if more dimensions within a song appear to be similar, then it should not be considered coincidence. However, these are subjective measures. Measures that may be routinized with the tests of time, but in no time soon.

When taken together in the context of music similarity, the conditions of the first two categories help prove the third. If there is the potential for inspiration, and there the probability for coincidence is low, then there is a high likelihood that plagiarism has occurred. This is when a song potentially violates copyright law. The prevalence of sampling in pop music today, makes this a particularly difficult and widespread issue in the Music Industry today. Where does inspiration start and stop? How different should a rhythm or melody be to be original? These are questions that the entire industry is grappling with.

3.2 Research Ethical Considerations

In the area bias as it relates to research, the questions are different. There are certain considerations we can be safe to consider moot. For example, the idea that music similarity research results could cause physical harm to listeners of the music is not plausible under normal circumstance. However, could this research be used to economically harm certain groups or individuals? Could it disadvantage an artist or group of artists as a result? Unfortunately, without proper controls and forethought the answer would be 'yes' there is a probability where benefits to some happen at the expense of others.

Unexpected or undue harm can occur as a result of "unintentional power and bias" dynamics in the research process [14]. Due to the complexities of algorithms, and the embedded nature of this intelligence into products, it creates a *black-box* problem we must be aware of. Holzapfel, et. al claim bias come in three forms:

- **Pre-existing** – relating to existing socio-cultural norms.
- **Technical** – relating to the data available, methods, or evaluation techniques
- **Emergent** – relating to how algorithms behave when faced with new *emergent* types of data that they have never encountered before.

As researchers, we would also posit that we have our own form of *research bias* that we must consider. We obviously want to help solve a problem, get good marks from our advisors, and be well known for the quality of our work. We have potential bias to create a favorable looking body of work, and are thus motivated to make choices about datasets, about problem scope, and about our technical approach that may inherently be biased. Each of these forms of bias have the propensity to potentially impact artists and other market participants.

To address these concerns, we carefully considered our research design, validating our choices along the way with our advisors. As guidelines we evaluated these decisions through the following lens:

Table 2. Bias Containment Challenge Questions

Bias Category	Challenge Questions
Pre-existing - Cultural	Are we considering the proper breadth of data to represent cultural minority classes, under-represented groups, and emergent categories?
Technical	Does our data choice contain a representative sample for the problem scope? Is the quality of the data such that we can be reasonably assured that it will not materially impact the results?
Emergent	Have we designed tests to validate results against new types of data properly?
Researcher Judgement	Do we have proper controls to document data or technical related issues to combat error propagation? Have we properly documented "ground truth" measures?

4 Survey of Techniques

Variety of prior research completed on detecting music similarities were focused on music style recognition, music similarity and genre categorization. Bogdanov, Serrà, Wack, Herrera and Serra [22] demonstrates exploring music similarity facilitating multi-media retrieval focusing on seeking a suitable distance measurement between songs based on predefined music feature space. Li and Ogihara [23] investigate acoustic based features to retrieve music features that focuses on similarity search and music sound emotion detection by applying Db8 wavelet filter and timbral features to generate compact music features. The general approach for similarity search is defined by Euclidean distance between two files and for emotion detection it uses decomposed multiclass classification which uses Support Vector Machine (SVM) to train on extracted features. In case of Logan and Salmon [24], who employs method of comparing songs solely based on their audio content which creates a signature for every song based on their K-means clustering of spectral features. From machine learning approach, Dannenberg, Thom, and Watson [25] illustrates how to build an effective style classifier by identifying 13 low-level features based on MIDI data classified using Bayesian, linear and neural network algorithms. Kuo, Shan, Chiang and Lee [26] observes a personalized content /emotion -based music filtering system and predicts the preference of new music by computing a weighted average of all ratings given by peer group of similar preference for end-user.

Our approach to the problem of finding music that is similar has been studied in past [22,23,24,25,26]. Where Logan and Salomon offer insight on the use of MFCC to define similarity [24], Bogdanov, Serrà, Wack, Herrera and Serra [22] exploits the rigor offered by Euclidian distance and SVM. Logan and Salmon [24] leverages K-means clustering to compare the underlying audio content and spectral features. However, we are exploring the spectrum of human-detectable pitches with various octaves which forms the DNA of musical features. We remained focus exploring *chroma* features to explore tonal and pitch content which essentially decomposes 12 pitch classes. Furthermore, capturing chroma musical characteristics in condensed and visual form to further process with Deep Neural Network to view low-level features of a song with reasonable feature encoding to target genre classification and similarity measurement.

5 Solution Approach

Given the scope of the project, it is critical to have a plan for each stage of the solution. This is important not only for the buildout towards analyzing the data, but it is also critical for fast iteration as new insights emerge that take research in new, interesting directions. It is with this in mind, and with the end goal of building reproducible research, that we outline our solution in this section.

5.1 Data Acquisition

As with any project of this type and ambition, it is critical to ensure we have enough high-quality raw data to work from. Further, this raw data requires high quality descriptive metadata to suite it. After reviewing various datastores for music data and metadata, the Free Music Archive (FMA) dataset was chosen [?]. The FMA dataset includes 106,574 tracks (songs) of royalty free songs in mp3 format, from 16,341 artists, along with metadata that covers artist names, album titles, 163 genres, artists biographies, and popularity metrics for each song. This archive represents roughly 1 Terabyte of untrimmed, raw, song data from which to create features, and provides a way for us to measure ground truth accuracy measures for modeling.

While the FMA archive includes a set of "common features," we chose to perform feature extraction on the data ourselves to a) maintain target control of the process, and b) to ensure reproducibility of our results on other data sets assuming similar metadata is available. As one might imagine, much hinges on feature generation when it comes to signal processing and handling raw data for these tasks.

Features were generated for each song, and then combined into a final, high dimension, feature repository. Total dimensionality of the processed feature space was 117. Categorization of these features is shown in the table below.

Category	Category Description	Dimensionality
Timbral	Tonal texture	20
Temporal	Time domain signals	15
Spectral	Musical characteristics by spectra	20
Rhythmic	Musical timing	12
Melodic	Melodic content	48

Commented [VC3]: All of these dimensions need adjusting for final count

As noted by Zhang, et. al., "high dimensionality of existing [low-level] audio features has restricted the applicability of content-based music retrieval in large collections" [8]. Later in this paper we discuss our application of dimensional reduction in an effort to simplify modeling procedures.

5.2 Exploratory Data Analysis

The FMA metadata includes several pieces of data to consider not the least of which is in genre and artist. While analyzing the data, there are some significant observations. The FMA dataset is highly imbalanced for each of these classes. This is true for both genres (n= 164) and artists (n=16881), but much less so for the latter. The compiled dataset consists of 59,800 rows and 8090 columns (see Figures below).

Most genres show up in the archive less than 3000 times, with a few genres having more than 23,500 songs associated to them. Therefore, any of 164 genres in the dataset that are less than 1% of the total row count, we will group them into an "other" column. Artists on the other hand have a frequency of occurrence ranging from at most 50 songs down to a single song, with most artists only being represented once or twice in the dataset. Both genre and artist name maintain no missing values in the data set. Also, every song has an assigned genre and artist.

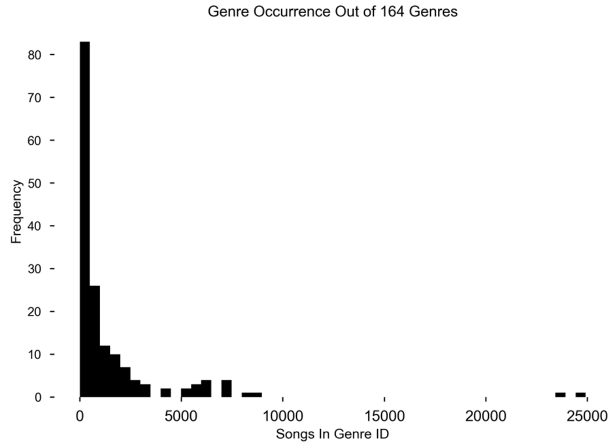

Fig. 5. Song Genre Class Imbalance

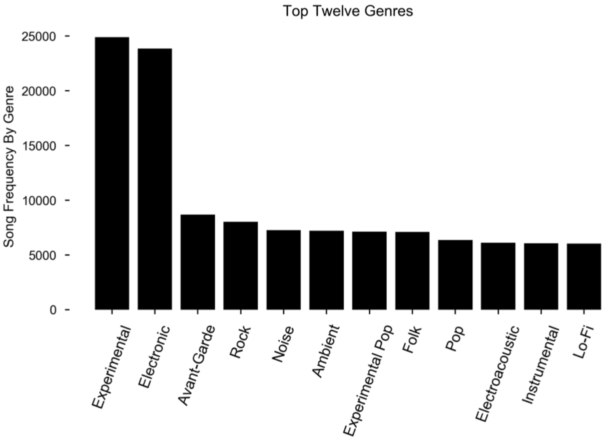

Fig. 6. Most Frequent Genres

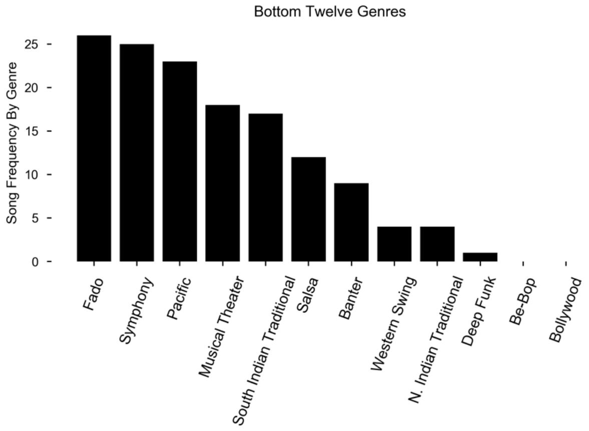

Fig. 7. Least Frequent Genres

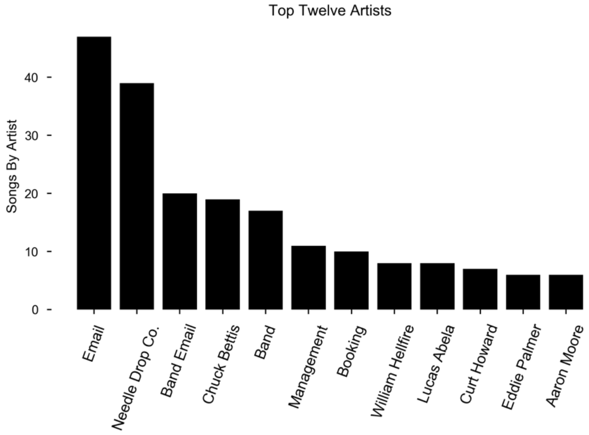

Fig. 8. Most Frequent Artists

5.3 Technical Approach

As mentioned previously, after acquiring the necessary data and metadata references, it is necessary to assemble a data processing pipeline to properly extract a feature space for each specified song. Processing such a large dataset for feature mining required careful planning and coordination, not only to properly perform the execution successfully on all files, but also in terms of properly managing the file system. This is where we look to the Cloud to help to solution our pipeline. We chose to utilize a publish-subscribe model to process each individual audio file, and then store the audio profile results (see Algorithm 1).

```
Algorithm 1:   Feature Extraction Algorithm
Input: Initial mp3 raw audio datafile;
Output: Unique JSON feature profile for the audio;
Description:
1. File is uploaded to Amazon Web Services (AWS) Simple Storage System (S3) bucket;
2. Creation event is triggered by file upload and a message digest is placed on the Simple Queue Service (SQS) message bus;
3. A batch process monitors and detects new message on the bus, and then pulls the message digest, including its unique identifier;
4. The batch process utilizes the digest to download the raw audio locally from S3;
5. The batch process then does feature extractions for the audio file and stores all of the results in a JSON file;
6. The batch process uploads the JSON file to S3;
7. Upon successful upload and verification of the JSON file, the batch process deletes the message off of the message queue;
8. The batch process cleans up the file system;
```

Once individual audio profiling is completed, we turn our attention to working with the feature space, and metadata. Given the feature space has XXX dimensions, it we anticipated performing feature reduction to better simplify the feature space. For this perform a principal component analysis (PCA) and to locate what components are required to still satisfy >90% of the variance in features. These YYY final components will be used for final modeling and analysis. In doing so, we are able to sacrifice very little feature information, while reducing overall dimensionality to ZZZ% of its original breadth.

As Purwin et. al. noted “neither within now across tasks is there a consensus on what input representation to use (log-mel spectrogram, constant-Q, raw audio) and what architecture to employ (Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) or both, 2D or 1D convolutions, small square or large rectangular filters), leaving numerous open question for further research.”[19] It is with this context where we propose to utilize a Deep Neural Network built with the training goal of being able to view low-level features of a song, and encode that song from those features. This approach allows us to take the encoder network and build necessary layers into it for genre classification and similarity measurement, as well as potentially artist recognition.

6 Analysis and Results

To extract useful information from large music collection data requires tools that are able to extract meaningful information related to music audio features like genre or style [27]. The music genre can help classify the most popular content description of any music file as illustrated as per Aucouturier and Pachet [28]. The pattern recognition techniques were employed to large volume of music data to generate musical features like beat, chroma, mel frequency cepstral coefficients (MFCCs) etc. We used Principal Component Analysis (PCA) to transform our high-dimensional feature representation into low-dimensional and balanced feature set while retaining the spectro-temporal feature set. The initial dimensionality reduction resulted in 150 principal components which were trained against genre classifier on an H_2O AutoML resulting in five potential models.

Model	mean_per_class_error	logloss	rmse	mse
XGBoost-1	0.715	3.407	0.901	0.812
XGBoost-2	0.717	3.330	0.903	0.815
XGBoost-3	0.712	3.522	0.893	0.799
XGBoost-4	0.721	3.337	0.914	0.835
GLM	0.730	3.434	0.937	0.878
Stacked Ensemble-1	0.730	3.434	0.937	0.878
Stacked Ensemble-2	0.730	3.434	0.937	0.878

7 Future Work

The future work content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularized in the 1960s with the release of Letraset sheets containing Lorem Ipsum

8 Lessons Learned

The lesson learned content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularized in the 1960s with the release of Letraset sheets containing Lorem Ipsum

9 Conclusion

The conclusion will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

References

1. Rentfrow, Peter J.; Goldberg, Lewis R.; Levitin, Daniel J. (1 January 2011). "The Structure of Musical Preferences: A Five-Factor Model" *Journal of Personality and Social Psychology*. 100 (6): 1139–1157.
2. Hardjono, Thomas, et al. "Towards an Open and Scalable Music Metadata Layer." arXiv preprint arXiv:1911.08278 (2019).
3. Müllensiefen, Daniel, and Marc Pendzich. "Court decisions on music plagiarism and the predictive value of similarity algorithms." Musicae Scientiae 13.1_suppl (2009): 257-295.
4. Dittmar, Christian, et al. "Audio forensics meets music information retrieval—a toolbox for inspection of music plagiarism." 2012 Proceedings of the 20th European signal processing conference (EUSIPCO). IEEE, 2012.
5. De Prisco, Roberto, et al. "Music plagiarism at a glance: metrics of similarity and visualizations." 2017 21st International Conference Information Visualisation (IV). IEEE, 2017.
6. Dredge, Stuart. “Music Industry Enters 2020 on a Wave of Growth – and Optimism.” Music Industry Enters 2020 on a Wave of Growth – and Optimism, Music Ally Ltd., Dec. 2019, musically.com/2020/01/03/analysis-music-industry-2020-growth/.
7. International Federation of the Phonographic Industry. IFPI Global Music Report 2019, 2 Apr. 2019, ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2019.
8. Zhang, Bingjun, et al. “CompositeMap.” Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 09, 2009, doi:10.1145/1571941.1572011.
9. Purwins, Hendrik, et al. "Deep learning for audio signal processing." IEEE Journal of Selected Topics in Signal Processing 13.2 (2019): 206-219.
10. Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).
11. “Music.” Wikipedia, Wikimedia Foundation, 9 Feb. 2020, en.wikipedia.org/wiki/Music.
12. Stav, Iyar. "Musical plagiarism: a true challenge for the copyright law." DePaul J. Art Tech. & Intell. Prop. L 25 (2014): 1.
13. “How to Tell If a Song's Been Copied - from a Trained Musicologist - BBC Newsbeat.” BBC News, BBC, 23 Sept. 2015, www.bbc.co.uk/newsbeat/article/34282895/how-to-tell-if-a-songs-been-copied---from-a-trained-musicologist.
14. Office, U.S. Copyright. “Legislative Developments.” Copyright, www.copyright.gov/legislation/dmca.pdf.
15. Shah, Ayush & Kattel, Manasi & Nepal, Araju & Shrestha, D.. “Chroma Feature Extraction.” (2019).
16. Holzapfel, Andre, Bob Sturm, and Mark Coeckelbergh. "Ethical dimensions of music information retrieval technology." Transactions of the International Society for Music Information Retrieval 1.1 (2018): 44-55.
17. G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, July 2002.
18. Tao Feng. “Deep Learning for music genre classification”

19. Purwins, Hendrik, et al. "Deep learning for audio signal processing." IEEE Journal of Selected Topics in Signal Processing 13.2 (2019): 206-219.
20. Mdeff. “Mdeff/Fma.” GitHub, 21 Feb. 2020, github.com/mdeff/fma.
21. Patch, Nick. "Meet the Man Classifying Every Genre of Music on Spotify–All 1,387 of Them." *The Toronto Star* (2016).
22. D. Bogdanov, J. Serra, N. Wack, P. Herrera and X. Serra, "Unifying Low-Level and High-Level Music Similarity Measures," in *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 687-701, Aug. 2011.
23. Tao Li and M. Ogihara, "Content-based music similarity search and emotion detection," 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Que., 2004, pp. V-705.
24. Logan, Beth, and Ariel Salomon. "A content-based music similarity function." Cambridge Research Labs-Tech Report(2001).
25. Dannenberg, Roger B., Belinda Thom, and David Watson. "A machine learning approach to musical style recognition." (1997).
26. Kuo, Fang-Fei, et al. "Emotion-based music recommendation by association discovery from film music." Proceedings of the 13th annual ACM international conference on Multimedia. 2005.
27. Panagakis, I., Benetos, E. and Kotropoulos, C., 2008. Music genre classification: A multilinear approach. In ISMIR (pp. 583-588).
28. Aucouturier, J.J. and Pachet, F., 2003. Representing musical genre: A state of the art. Journal of New Music Research, 32(1), pp.83-93.