UMAP Enhanced Clustering for Music Genre Classification

Ricardo Inácio
up202302742@up.pt
Faculdade de Engenharia Universidade do Porto
Porto, Portugal

Fábio Morais
up202008052@up.pt
Faculdade de Engenharia Universidade do Porto
Porto, Portugal

António Oliveira
up202008004@up.pt
Faculdade de Engenharia Universidade do Porto
Porto, Portugal

Pedro Gomes
up202006086@up.pt
Faculdade de Engenharia Universidade do Porto
Porto, Portugal

Abstract

Since the widespread adoption of streaming platforms, an unfathomable amount of music was made readily available to the average consumer. Consequently, the need for robust recommender systems arose, so that each user is able to find the contents that mostly resonate with their tastes. To build them, characterising features about the data need to be promptly available, so that records can be properly grouped and organised, based on diverse factors of interest. Given that this is not usually the case, due to intellectual property restrictions or cost, several methods have been developed across the years, as a way to systematically capture patterns on music data, to successfully classify and cluster them. This makes music naturally comparable, thus easier to recommend. The state of the art of such technologies comprises several deep learning methods, which although effective, usually imply high computational costs, and a lack of interpretability. To that end, we propose a simple method based on supervised uniform manifold approximation and projection, to create a three-dimensional latent space, based on features extracted from song files, to which one can use to classify new samples, transparently. Different strategies are compared, based on distinctive guided clustering approaches (supervised and pre-clustering), to assess which yields optimal results. Given the simplicity of the employed approaches, the computational cost is low, and the decisions made can be easily interpretable with XAI methods tailored to tabular data. Experiments show that this approach is effective in classifying novel, out-of-distribution samples, irrespective of the apparent simplicity of the system. The code and detailed analyses are available online[1].

CCS Concepts

• **Computing methodologies** → *Learning latent representations.*

Keywords

Music Classification, Clustering, Latent Representation Learning, Dimensionality Reduction

1 Introduction

As streaming platforms became the *de facto* avenue to listen to music for the majority of the population, an unprecedented quantity of music, of diverse genres and subgenres, was rendered easily accessible. Thus, to ensure that one can navigate through virtually limitless possibilities, and reach serendipitous outcomes, robust recommender systems became essential for the optimal execution of such services. To that end, a need for large quantities of musical data, containing the necessary characteristics for intuitive categorisation, and subsequent grouping, arose. Modern techniques based on deep learning, such as convolutional neural networks (CNN), have reached high efficacy, while requiring minimal manual effort, by learning from raw audio data directly. Nonetheless, such methods require vast amounts of high quality data, to learn the large numbers of parameters scattered across the several layers of the network. These also imply large computational costs to effectively train, and sometimes simply to use with reasonable performance.

Classical machine learning approaches, although requiring data preprocessing and feature engineering steps, are still commonly applied based on the simplicity of their implementation, and low computational cost. These may not, however, be able to reach the levels of accuracy deep methods attain, for more complex and nuanced tasks. Yet, if the goal is to simply organize different records regarding the genre each belongs to, and to clearly interpret how certain characteristics leads to a specific categorisation, simpler methods bay be a better fit.

To address this, we propose a method based on supervised uniform manifold approximation and projection, to reduce a set of musical features, which can be automatically extracted by frameworks such as `librosa` [6] and `TIV` [8], into three dimensions, to then semantically place the songs. As the resulting latent space is so low-dimensional, the data points can be easily plotted in three-dimensional visualizations, hence increasing the explainability factor of the method. Not only the features from novel records can be analysed and interpreted, but also clearly visualised in the 3D space, which helps in discerning which other songs are the most similar. Then, a more comprehensive analysis can be conducted to uncover how songs relate to each other. The goal of this work is not to obtain a particularly precise classifier, but to develop a method that automatically extracts musical features from a given song, to then analyse how the respective characteristics place it within the laten space, and how similar it is to other songs, via simplistic approaches.

We experiment with different configurations to obtain the best latent representation for all the genres. We employ a supervised approach, which uses the target labels of the chosen dataset, to improve the separability of each cluster. We also experiment with an unsupervised k-means-based approach, since the number of classes is known beforehand, to verify if it leads to improvements by not considering the intrinsic structures of the learned data.

[1] https://github.com/ricardoinaciopt/seminars3b

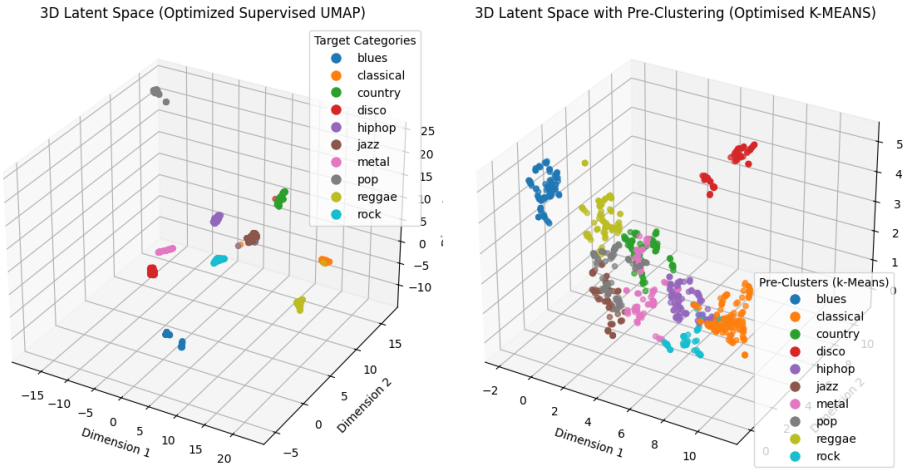

Figure 1: Latent space representations of both employed approaches. In the left, the space was constructed using target labels in a supervised manner. In the right, it was constructed using k-means to pre-cluster the data by inferring labels.

2 Dataset

We employ the GTZAN "genres original" dataset in the experiments [11], as it is one of the most prevalent choices in the literature of machine learning algorithms applied to musical and audio classification or information retrieval. This specific collection encompasses 10 genres of music, each containing 100 songs with 30 seconds of length. Being considered the "MNIST of Audio", this benchmark dataset is extremely useful to evaluate the performance of our approach, in relation to more efficient solutions in the field.

To ensure the proposed methodology remains as broad as possible, we applied our feature extraction pipeline to ensure any valid audio file is represented equally when given as input to our system. So, for each song in the collection, several features were extracted, and placed in a row of the final dataset, having the name of the corresponding folder, which represents the genre of the song, as the target feature. This led to a well-balanced dataset, that can be used in a multi-class classification task, being the goal to relate a new record with one of the classes, or genres.

3 Feature Representation

To build the feature set, we leveraged two automatic extraction packages: `librosa` and `TIV`. A total of 20 features were extracted: 16 from the first, and 4 from the second. The subset acquired from `librosa` is scattered along *spectral*, which define the frequency of the signal, *temporal*, defining features in the audio domain, *mel-frequency cepstral coefficients (MFCCs)*, derived from the mel scale, *harmonic* and *percussive* features, capturing energy, rhythm, and timbre. From `TIV`, Chromaticity, diatonicity, dissonance, magnitude are collected. These reveal glimpses into the harmonic and tonal characteristics of the records. These are then properly structured in a tabular format, each value in a column associated with the relative feature, in a single row for every song in a dataset. This ensures compatibility with traditional machine learning algorithms, and can serve as input to the dimensionality reduction algorithm, to project the dataset into the latent space, in three dimensions.

3.1 Spectral Features

- **Spectral Centroid** - Indicates the "center of gravity" of the spectrum. It is used as an indication of brightness being commonly used in music analysis and genre classification [2].
- **Spectral Bandwidth** - Width of the spectrum around the centroid, it may be used to distinguish clean tones from noisier ones.
- **Spectral Rolloff** - Measures the bandwidth of the audio signal by determining the frequency bin under a given percentage of the total existing energy. In this project, librosas's default 85% was used.
- **Spectral Contrast** - Compares the intensity difference between peaks and valleys in the spectrum. High contrast values correspond, generally, to clear, narrow-band signals while low contrast values correspond to broad-band noise [4].

3.2 Energy and Temporal features

- **RMS Energy** - Root-Mean-Square value for each frame of the audio.
- **Zero Crossing Rate** - Rate at which the signal crosses the zero amplitude mark. As it evaluates the possible times the sound goes from positive to negative and vice-versa, it gives a representation of the smoothness of the wave.
- **Tempo** - The estimated pace of the music is measured in beats per minute.
- **Mean Beat Interval** - Average time duration between consecutive beats.
- **Onset Strength Mean** - Average intensity of note onsets (moments when notes begin) throughout the track.
- **Onset Strength Variance** - Variation in the intensity of note onsets over time.

3.3 Mel-frequency Cepstral Coefficients(MFCCs)

MFCCs are a widely used set of features in speech and music analysis that help capture the timbre or tonal quality of an audio signal. Derived from the Mel scale, which is a perceptual scale of pitches that better reflects human auditory perception, MFCCs are designed to approximate how humans hear sound. This project used the following features:

- **MFCC Mean** - Average values of the MFCCs across the audio signal.
- **MFCC Variance** - Variation in the MFCC values.

3.4 Harmonic Features

- **Harmony Mean** - Average energy at harmonics of time-frequency representation
- **Harmony Variance** - Variation in the energy at harmonics of time-frequency representation

3.5 Percussive Features

- **Percussive Mean** - Average intensity of the percussive component.
- **Percussive Variance** - Variation of intensity of the percussive component.

3.6 TIV.LIB Features

These features were generated using the tiv.lib. We refer the reader to the work by Ramires et. al. [8] for a more detailed description.

- **Chromaticity** - Measures the concentration of a sonority within the chromatic pitch circle, providing a value between 0 and 1.
- **Diatonicity** - Measures the concentration of a sonority within the circle of fifths, producing a value between 0 and 1.
- **Dissonance** - measures dissonance by subtracting the normalized TIV magnitude from 1. It uses weighted TIV coefficients to rank intervals based on empirical dissonance ratings.
- **Magnitude** - is a 6-element vector that characterizes the harmonic quality of a pitch profile by reporting the magnitudes of its TIV components, reflecting intervallic content and tonal qualities, and remaining invariant under transposition or inversion.

4 Dimensionality Reduction Algorithm

In order to convert the tabular dataset of extracted features, into a three-dimensional latent space, the `UMAP` dimensionality reduction technique. This method is advantageous as it is particularly good at preserving both local and global structures of the original data [1], which is important in more complex concepts such as musical genres. This approach builds a high-dimensional graph representation of the original data, to then find a similar, low-dimensional graph that resembles the same conceptual structures [12].

In the proposed method, two different approaches are employed: supervised class-based clustering, and pre-clustering. In the first, the categoric target labels, which denote the genre of each sample in the GTZAN dataset, are directly used to aid the reduction algorithm to build more tightly coupled clusters, regarding the groups represented in the data. The second, employs a clustering algorithm (e.g., k-means) to find prospective clusters, which are simply used to group the data points in the latent space. It should be noted that, in the first, the target labels are directly applied while constructing the latent data space, while in the second, the pre-computed clusters are simply used after building the latent space, to semantically separate the points. The first strategy leads to more closely bonded songs, in highly separated clusters, and the second disperses all points in a chaotic mass, having each group somewhat close-together, with some genres overlapping others. Both methodologies are disclosed in Figure 1, exhibiting the readily apparent distinct characteristics.

Any clustering algorithm could be theoretically used to infer the class labels in the pre-clustering strategy, however, we used k-means in the experiments. This algorithm was selected, given it has been shown to perform remarkably well with normalised reduced dimensional spaces in the literature [3], particularly when paired with UMAP reductions. Also, besides high achievable performance, this method is very efficient when working in lower dimensions [10], as it is the case with our 3D space. It should be noted that the performance of the algorithm decreases if the number of classes (in this case, genres) is too large [5], thus, for a simple 10-class task, it was a suitable choice.

In a general clustering context, the first approach would be the best, since it leads to more well-defined clusters. However, in the topic of musical genre classification, the prospect of overlapping groups would perhaps lead to more satisfactory results, as by nature, genres may share characteristics that cause them to blend with each other. A comparison of both methods is detailed in Section 6.

5 Methodology

The workflow behind the proposed approach comprises development and inference stages, each with three main steps: i) feature extraction, ii) dimensionality reduction, iii) projection into latent space. In the development phase, the full audio collection from GTZAN, is converted into a tabular dataset of the extracted features described in Section 3, and passed as input to the dimensionality reduction algorithm `UMAP`, after normalizing the values using a *StandardScaler*, to convert it into a three-dimensional space. Each point of the dataset is then placed within that space. In the visualization, each point is coloured based on the target label, either passed in a supervised manner, or inferred by a clustering algorithm (e.g., k-means). In the inference stage, a new song is passed as input to the pipeline, which is immediately clipped into a fixed length of 30 seconds, if it is longer than that, to ensure consistency with the rest of the dataset. Then, features are extracted, and projected into the 3D latent space using both `UMAP` approaches, to then be able to compare the performance of each, in correctly placing a new data point near similar ones.

Using a simple implementation of the nearest neighbours (KNN) algorithm, based on the `numpy` Python package, we can find the most likely genre of a new sample based on its latent representation. Given the latent space constructed from the reduced and normalised tabular dataset, alongside the reduced and normalised

representation of a new song, we can calculate its Euclidean distance to all the previously learnt songs. Then, we sort the distances in ascending order and select the 150 closest songs, to find their genre, assigning to the new song the most recurrent one (which should be placed at the top of the ordered list). This high number of neighbours is usually not necessary, and it was simply chosen for the sake of interpretability, to analyse where each new song was being placed relatively to other samples in a global sense.

5.1 Feature Importance

In order to have a meaningful evaluation of what distinguishes each genre from more than a purely latent space perspective, we must understand what features of our dataset are most relevant for such an analysis. With this purpose in mind, a selection process was conducted using a random forest algorithm and mutual information classifier. The first was utilized for its *feature_importances* parameter which returns the impurity-based feature importances used to build the random forest, while the latter is based on the notion that between two random variables we are able to get a value that represents the amount of information we are able to get about one while observing the other. The values from each of these methods were averaged out, and the top 10 results were selected for analysis, as we present in the following subsections.

Feature	Importance
Onset Strength Mean	0.307119
Percussive Variance	0.299668
Spectral Rolloff	0.247631
Spectral Bandwidth	0.241894
Spectral Centroid	0.235849
MFCC Variance	0.234867
Onset Strength Variance	0.229139
MFCC Mean	0.226322
RMS Energy	0.217879
Dissonance	0.214118

Table 1: Top 10 features from the feature importance analysis

Firstly we created a correlation matrix for these 10 features, in order to identify potential interactions between them.

From Figure 2, we can see that the spectral features are highly correlated amongst themselves, with onset strength variance and mean, being naturally correlated as well. Interestingly MFCC variance seems to have very little correlation with any other features apart from its mean counterpart.

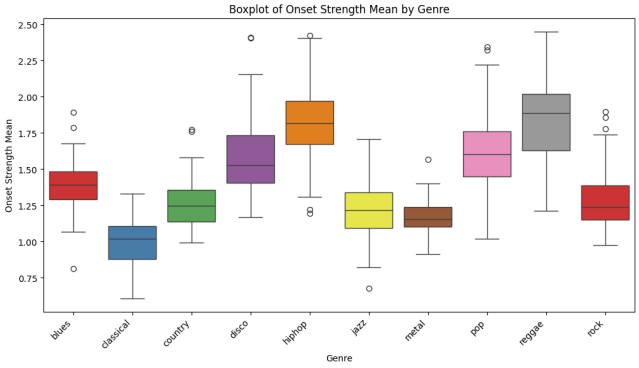

Figure 3: Boxplot of Onset Strength Mean by Genre

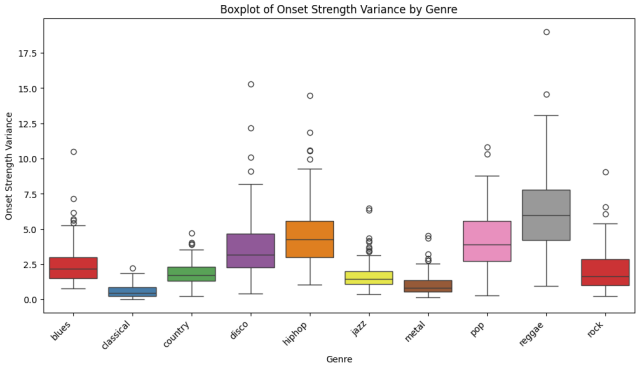

Figure 4: Boxplot of Onset Strength Variance by Genre

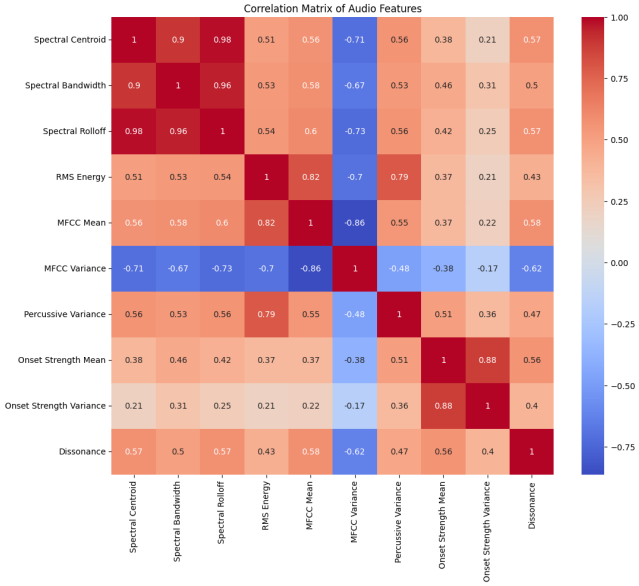

Figure 2: Correlation matrix of all features in the dataset

5.1.1 Onset Strength. Genres like disco, hip hop, and reggae are similar in this metric, with their very pronounced rhythmic beats leading to big jumps in the intensity of the notes being played, while classical music features more mellow and progressive transitions and starts registering lower values. It was still surprising that metal registered a relatively low mean and small variance for its onset strength, something that we would expect to be characteristic of this genre; this could most likely be attributed to the fact that there aren't usually big variations in note "strength".

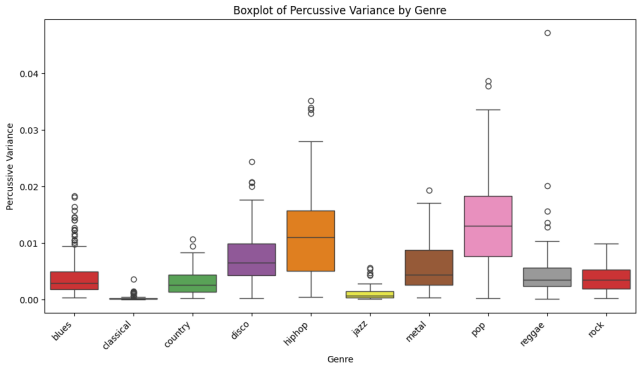

Figure 5: Boxplot of Percussive Variance by Genre

5.1.2 Percussive. Similar to onset strength, genres that feature heavy use of percussive elements, such as hip hop, disco, and pop end up having a pretty high percussive variance, as expected. Notably metal's heavy drums and tempo-keeping instruments shine through in this metric, granting it a decently high value.

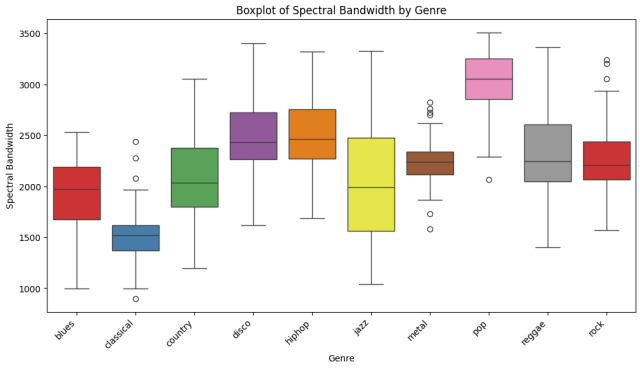

Figure 6: Boxplot of Spectral Bandwidth by Genre

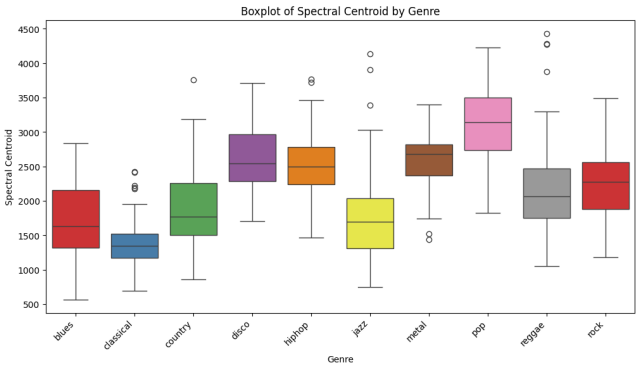

Figure 7: Boxplot of Spectral Centroid by Genre

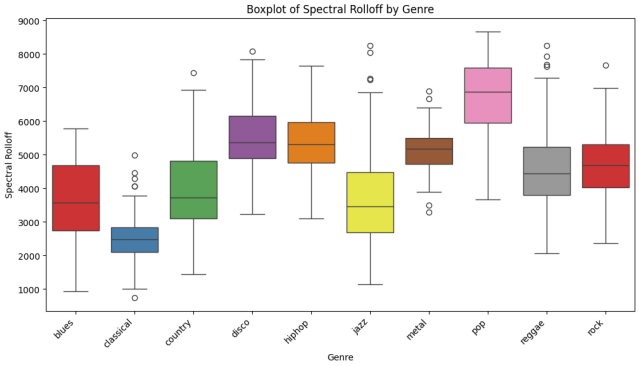

Figure 8: Boxplot of Spectral Rolloff by Genre

5.1.3 Spectral. As demonstrated in the correlation matrix (Figure 2), all 3 spectral features in this top 10 are very similar, and since one of them is important it is to be expected that all other ones would follow suit and be of equal, or similar, importance, with the exception of spectral contrast, which is not as closely correlated as rolloff, bandwidth, and centroid.

From the results, we can see that brighter-sounding genres, such as pop, end up with larger values for these metrics, and all genres share a big range of values within them. Jazz's large range of different subgenres, instruments, and influences is also highlighted. Unexpectedly, classical music does not register that high in spectral metrics, which could be attributed to the samples of the dataset that we are using, explaining the larger number of outliers, samples that may have sounds that register higher.

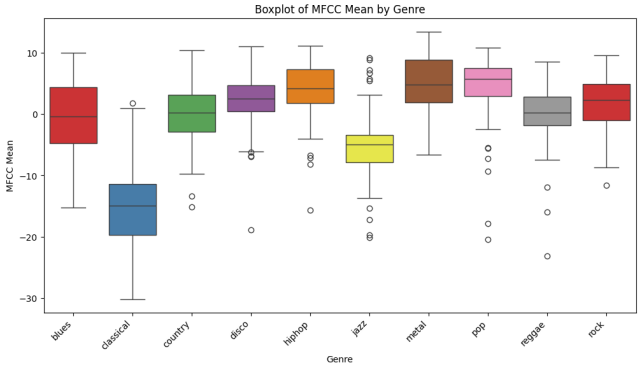

Figure 9: Boxplot of MFCC Mean by Genre

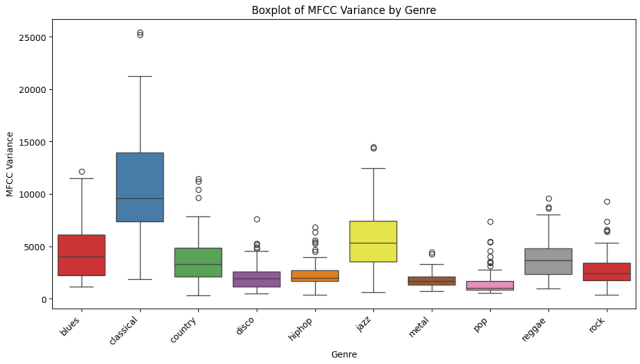

Figure 10: Boxplot of MFCC Variance by Genre

5.1.4 MFCC. Genres like blues, country, disco, hip hop, metal, and pop share a similar tonal quality, as to be expected given their cross-influences and the large number of subgenres and fusions that inevitably overlap. Interestingly blues, reggae, and country feature the most dispersed set of values, something that despite unexpected may be attributed to the samples utilized.

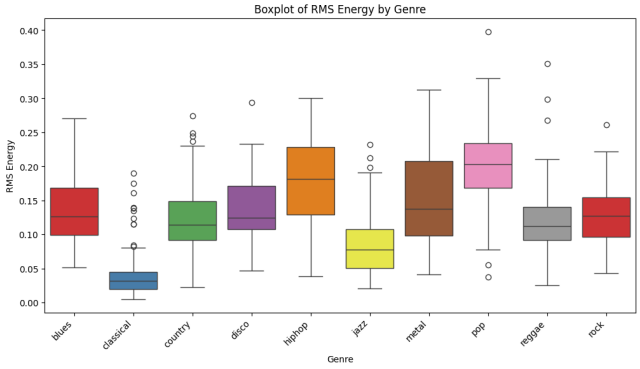

Figure 11: Boxplot of RMS Energy by Genre

5.1.5 RMS. RMS energy is a difficult feature to analyze and distill to a simple interpretation, given its analytic and mathematical nature, however, it may be interpreted as a measurement of the loudness of an audio segment, with this in mind we can see that the results for metal, hip hop, and pop are to be expected, with most of the other genres hovering around the same levels.

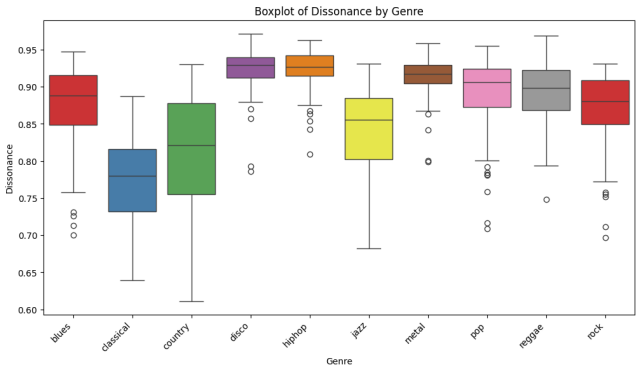

Figure 12: Boxplot of Dissonance by Genre

5.1.6 Dissonance. As expected classical, country, and jazz, featuring more harmonious notes, end up with lower dissonance values, whereas more recent genres such as disco, hip hop, and metal, with their clashing, harsher notes and more layered noises place higher on this metric.

6 Experimental Results

To provide a preliminary assessment of how each employed method constructs the genre-based clusters, the *silhouette* metric was chosen. This metric takes into consideration the intra-cluster distance each point has to all other points in the same cluster, while also measuring the inter-cluster distance to all other points of a neighbouring cluster. A positive value means a point is assigned to the correct cluster, a negative value means the opposite, and a near-zero value indicates ambiguous results, likely placed in the inter-class boundary [9]. We employ the `scikit-learn` implementation of the algorithm, to which we pass as parameters the *latent space* alongside the real *encoded targets*, in the supervised method, and the *latent space* alongside the *cluster labels* inferred by the employed *k-means* algorithm, in the pre-clustering one. With the goal of obtaining the optimal set of parameters in both approaches, we used *optuna* to search the parameter space, for 20 trials each. In the end, the *silhouette* score of the supervised approached attained 0.88, while the pre-clustering one achieved only 0.52. This was expected, simply by visually analysing the resulting landscapes (c.f. Figure 1), as the pre-clustering approaches showcased clusters that were not very well-defined, with some degree of overlapping. However, as previously stated, this may not indicate a worse latent representation given the nature of the task.

The experimental setup consisted in simply acquiring 5 random songs, from 5 random genres, available in the *Pixabay* database [7], to then pass as input to the devised pipeline. In the end, the results of the classification, and the extracted features were recorded.

Therefore, by examining Table 2, it is possible to assess how, even though the pre-clustering approach achieved a lower *silhouette* score, implying the clusters are not as well-defined as the ones from the other approach, it led to a higher accuracy while classifying novel songs. This is likely related to the nature of the task, since even though a song may have a specific genre assigned, the underlying characteristics of said genre may be common to several others, thus, a latent space where some overlap exists may be beneficial.

New Sample Genre	Supervised	Pre-clustered
Metal	Jazz	**Metal**
Country	**Country**	Pop
Classical	Country	**Classical**
Hip-hop	Jazz	**Hip-hop**
Disco	**Disco**	Reggae
Silhouette	*0.88*	*0.52*

Table 2: Comparison of the experimental classification results. Bold font denotes a correct classification. The pre-clustered approach, although attained a lower *silhouette* score, showcased greater accuracy. The *silhouette* score each method achieved is also shown, as reference.

Realistically, a song would hardly contain elements pertaining to a single genre, so the fact that a large number of neighbours is taken into consideration, may help in deciding on several related categories. In contrast, the supervised approach, although it leads to more precise clusters, could potentially be overfitting to the dataset, resulting in assessments that may be too strict for music from a different distribution, as new songs ought to surely fall between the defined classes.

From the analysis, it seems that the pre-clustering approach seems a better fit for music that blends elements of different genres, such has *Metal* and *Hip-Hop*, or for genres whose characteristics comprise ensembles of multiple, complex elements, such as *Classical*. The supervised approach, seems more appropriate for music genres that have well-defined characteristics, and can be easily discerned among others, such as *Disco*. The misclassification of the supervised approach indicate a bias towards music with string components, long solo instrument sections, and specific mellow vocal techniques, such as *Jazz* and *Country*.

In summary, although the clusters originating from the supervised approach are technically superior, as they are more closely-defined to the ground-truths of class belonging, as shown by a greater *silhouette* score, this can make it harder to classify novel samples. For in-distribution data, this approach would assign each sample the correct genre with high certainty, however, for out-of-distribution samples, since the underlying characteristics may be completely distinct, wrong genres might be assigned based on the poor reasoning, as it was seen in the experiments, where the method was biased towards *Jazz* and *Country*. For the pre-clustering method, even though clusters were poorly defined, with considerable overlap, for the given task this fact might be beneficial, as the concept of genre belonging may be somewhat ambiguous, thus making it more universal. This is the reason for increased performance, even with lower silhouette scores. We concluded that each method has its benefits and issues, that should be taken in consideration when applying to novel instances, such as audibly and easily discernable characteristics that might be indicative of class.

7 Conclusion

Although the genre of a song is rather difficult to define, given the inherent ambiguity of the concept, and the large number of different factors, such as musical, social, and historical, that contribute into properly characterizing it, we are still able to achieve some semblance of a computational characterization of genre. We achieve this solely based on mathematical and musical concepts alongside a strong computational foundation of simple machine learning methods and algorithms. The results, albeit interesting, we were able to achieve with this methodology, reinforce the previous notion. Being able to visualize the 3D latent space between the songs, labelled by genre, of our dataset, was an important step to understand the complicated relationships between each, and novel records. The conducted analysis allowed us to understand what such methods rely upon, to actively interpret sets of features, as vessels to represent specific musical genres. The code and detailed analyses are available online[2].

References

[1] Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vázquez-Baeza, Y., Mishne, G., and Knight, R. Uniform manifold approximation and projection (umap) reveals composite patterns and resolves visualization artifacts in microbiome data. msystems, e0069121, 2021.

[2] Grey, J. M., and Gordon, J. W. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America 63*, 5 (1978), 1493–1500.

[3] Hozumi, Y., Wang, R., Yin, C., and Wei, G.-W. Umap-assisted k-means clustering of large-scale sars-cov-2 mutation datasets. *Computers in biology and medicine 131* (2021), 104264.

[4] Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H., and Cai, L.-H. Music type classification by spectral contrast feature. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on* (2002), vol. 1, IEEE, pp. 113–116.

[5] Kim, K., Yun, W., and Kim, R. Clustering music by genres using supervised and unsupervised algorithms. Tech. rep., Technical report, Stanford University, 2015.

[6] McFee, B., Lostanlen, V., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Lee, K., et al. librosa/librosa: 0.8. 0. *Version 0.8. 0, Zenodo, doi 10* (2020).

[7] Pixabay. Royalty free music downloads. https://pixabay.com/music/search/. Accessed 30-12-2024.

[8] Ramires, A., Bernardes, G., Davies, M. E., and Serra, X. Tiv. lib: an open-source library for the tonal description of musical audio. *arXiv preprint arXiv:2008.11529* (2020).

[9] Shahapure, K. R., and Nicholas, C. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)* (2020), IEEE, pp. 747–748.

[10] Stern, S. Analysis of music genre clustering algorithms. Master's thesis, The University of Wisconsin-Milwaukee, 2021.

[11] Sturm, B. L. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461* (2013).

[12] Trozzi, F., Wang, X., and Tao, P. Umap as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: A comparison study. *The Journal of Physical Chemistry B 125*, 19 (2021), 5022–5034.

[2] c.f. footnote 1