Computing Science

Music Subgenre Classification Using Deep-Learning and Traditional Machine-Learning Approaches

Hariss Ali Gills

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: December 5, 2024

CS4040 Report

Department of Computing Science
University of Aberdeen
King's College
Aberdeen AB24 3UE

December 2024

Music Subgenre Classification Using Deep-Learning and Traditional Machine-Learning Approaches

Hariss Ali Gills

1 Introduction

Music has made a dramatic shift to digital platforms. As a result of streaming platforms' availability and ease of use, Spotify has grown significantly,increasing its user base from 15 million to 100 million since 2010 [3]. With over 60,000 tracks are now uploaded to Spotify every day [8], there are no doubts that automating the classification of music into genres or subgenres would be favorable.

Genre is a subset of music classified by a particular style, but a genre can also be guided via culture, business, particular artists, and other musical influences [4]. Genres can be further divided into subgenres. Examples include Country, Pop, Jazz, Rhythm and Blues, Metal, and Rap. It is evident, though, that music subgenres are discussed in less exact terms, wherein not all of the traits of the parent genre are always present in a particular "subgenre"; on the other hand, a subgenre may display traits that are not present in its parent [7]. For instance, some subgenres of Pop are Dance-pop, K-pop, Pop rock, Bubblegum pop, Europop, Latin pop, Country pop, Power pop, and Indie pop.

While there has been a large interest in classification on music into genres, subgenre classification has received attention only in EDM [1], likely due to lack of datasets other than GZTAN [13]. The term "Electronic/Dance Music" to refers to a diverse range of music composed using computers and electronic instruments, frequently for dancing purposes [12]. Metal is another captivating genre that has overlapping musical elements and subjectiveness since fans and critics often disagree on what defines a subgenre, as some bands can fit into multiple categories [18]. Hence, this paper, aims to find how accurate machine learning classifiers are within metal subgenres. Additionally, it can be asked if there is any significant confusion among subgenres over multiple classifiers.

2 Background and Related Work

As previously alluded, genre classification has been significantly more studied than subgenre classification. Work by Caparrini et al. investigated the automatic classification of different taxonomies for EDM music [1]. This paper interestingly points out that folk music has been analyzed for its cultural context. The authors used Beatport's 120 second previews to create two datasets that were fed to four models, namely Decision Tree, Random Forest, Extremely Randomized Trees, Gradient Tree Boosting via 10-fold cross-validation with 92 input variables. Two datasets were used since the taxonomy of EDM music changes rapidly. The study showed that the gradient tree boosting classifier outperformed the very randomized trees technique in Set 2 (accuracy of 48.2%), while the gradient tree boosting classifier outperformed Set 1 (accuracy of 59%) based on the mean accuracy values. The

confusion matrices, which used the best performing classifier per genre, showed that the subgenres are significantly confused asymmetrically.

Another study from 2011, also investigates classification of metal music [17]. Four classifiers were tested using a dataset that included 210 recordings from seven different subgenres. A custom classifier algorithm classified 37.1% of test samples correctly, which is significantly better performance than random classification (14.3%). k-NN found an accuracy of 42.8%, and the best result with correct classification rate of 45.7% was achieved by AdaBoost. Interestingly, when the number of subgenres was increased to 17, the best results were between 8-10%.

An intriguing approach taken by Ndou et al. concluded that 3 second duration input features can provide better accuracy than 30 second duration input features [13] when classifying genres using the popular GZTAN dataset. *Phase A*, *Phase B*, and *Phase C* were the three stages in which this study was carried out. The first two phases varied the input dimensions (from 51 to 223) on thirty second snippets fed to Linear Logistic Regression, Random Forest, Support Vector Machines, Multilayer Perceptron, k-Nearest Neighbour, and Naíve Bayes models. Naíve Bayes was dropped in the last phase, presumingly due to the low accuracy, for a Deep Learning approach using Convolutional Neural Network (CNN) with three second snippets. As for the results, the *Phase C* k-Nearest Neighbours model provided the best accuracy at 92.69%. All of the other models were more accurate in *Phase C* other than Logistic Regression and Support Vector Machines, which were most accurate in *Phases A* and *B* respectively. Only one confusion matrix was provided, the Logistic Regression during Phase A showed that the confusion was more symmetric. Ten country music excerpts were classified as rock music whilst being the most misclassified genre.

The above work was informed by upon Deep Learning research done by Pelchat et al. that used spectrograms 2.56 seconds of the song only into a CNN to classify the songs into genres [15]. Additionally, the work by Ndou et al. takes up the suggestion of using all the slices for a song instead of one per song and using the ReLU activation function. After some modifications to the number of layers and number of genres, an accuracy of 85% was achieved. This is higher compared to 66.50% found in Phase C of the previous study, likely due to the differences in datasets.

3 Research Question

As seen in the previous section, works that classify music into genres, often use the same dataset - GTZAN. Although results obtained via GTZAN are still meaningful, GTZAN has issues like repetitions, mislabelings, and distortions [16]. An attempt to classify subgenres has been made, but it uses high dimensional inputs which lead to longer training times. This report attempts to identify the most accurate Machine Learning model to classify subgenres in a genre

The research questions are as follows:

1. How accurate are machine learning and deep learning classifiers within 18 Metal subgenres?

2. Is there any significant confusion among subgenres over multiple classifiers? Is this a potential cause of redundancy in the subgenre?

Since the k-NN had the highest accuracy in previous works, it is expected to be the answer to the first question [13]. As for the second question, it's highly likely that there will be confusion because of work by [1], and the intuitive overlap caused by subgenres evolving from one another.

The hypotheses naturally flow as follows:

1. k-NN is the most accurate machine learning model to classify 18 Metal subgenres.
2. There is more than 40% confusion between two subgenres over at least two classifiers. This will be done qualitatively.

3.1 Models

In order to answer the above questions, an experiment will be conducted that adopts a similar strategy as *Phase C* in the work done by Ndou et al., but with a different dataset [13]. We utilize Linear Logistic Regression, Random Forest, Support Vector Machines, Multilayer Perceptron, k-Nearest Neighbour, and Naíve Bayes models from the `scikit-learn library` [14]. For the deep learning approach, the Convolutional Neural Network (CNN) model architecture is reused. This model includes an input layer followed by five convolutional blocks with a ReLU activation function [6].

3.2 Audio Features

`librosa` is a common and powerful tool to analyse audio [11]. The mean and variance of the following types of features are extracted:

- Magnitude-based: Represent timbral qualities like loudness, pitch, and compactness.
- Tempo-base: Capture rhythm and tempo characteristics, such as beats per minute and audio signal intensity.
- Pitch-based: Describe pitch aspects, contributing to harmony, key, and melody.
- Chordal progression features: Examine pitch chroma, representing pitch classes in a twelve-dimensional vector.

3.3 Dataset

Many sources were considered for gathering metal tracks, but the Spotify Web API stood out as the most flexible and representative of what metal fans listen to compared to other platforms. While the Free Music Archive does offer some metal tracks, it does not truly capture the broader range of what fans of the genre enjoy. To create a comprehensive dataset using python, spotipy, which is a lightweight Python library for the Spotify Web API, is utilized. Unlike Beatport, Spotify offers 30-second previews of music hosted on its platform, making it a more valuable resource for accessing popular metal tracks from the subgenres listed on "The Metal Archive" [2]. Unfortunately, the API only provides genre per artist not by track. Playlists titled "`Subgenre` Mix" playlists need to be

avoided since they are curated on per-user basis. So for each subgenre, the most popular public Spotify playlist with at least 100 tracks is used.

3.4 Accuracy

Since the tracks in each subgenre are of equal lengths, the amount of 3 second slices per subgenre should be balanced. However, some tracks might be unavailable to preview. Hence, the F1 score is also calculated [9].

4 Experimental Design

The null hypothesis for this experiment are:

1. k-NN is not the most accurate machine learning model to classify 18 Metal subgenres.
2. There is not more than 40% confusion between two subgenres over at least two classifiers.

To test for the hypotheses, the same models, hyperparameters and features will be used as *Phase C* by Ndou et al. [13]. A new dataset is created for this experiment. This dataset contains 1800 tracks. For each subgenre, the most popular public Spotify playlist with at least 100 tracks found on 2024/11/20 is used. After compiling the tracks, the features as described by Table 2 are extracted. For each track, 9 slices are taken. This is because `librosa` calculates the last slice to be around 2.7 seconds so it is discarded for being too short. The tracks were not downloaded to disk due to copyright and ethical concerns. Instead, Python's `io.BytesIO` handles them directly in memory. Additionally, Spotify rate limits the API, hence the feature extraction process had to be split into two spaced apart sessions. One final deterrent is that some tracks are unavailable to the UK market [5] so the final dataset has 14,572 slices.

Classifier	**Hyperparameters**
k-Nearest Neighbours	nearest neighbours=1
Multilayer Perceptron	activation=ReLu, solver=lbfgs
Random Forests	number of trees=1000, max depth=10, $\alpha = e^{-5}$, hidden layer sizes=(5000,10)
Support Vector Machines	decision function shape=ovo
Logistic Regression	penalty=12, multi class=multinomial

Table 1: Classifier Hyperparameters

Each of the above classifiers were evaluated using stratified 3-repeated 10-fold cross validation. As per scikit-learn's documentation, "Stratified k-fold is a variation of k-fold which returns stratified folds: each set contains approximately the same percentage of samples of each target class as the complete set." This results in 30 runs per classifier with the same train and test splits, in which the accuracy and F1 scores are calculated. For the second hypothesis, the confusion of each classifier is calculated as a mean due to the stratified nature of the folds.

For statistical significance, Shapiro-Wilk test is done on the accuracies to check for normal distribu-

Features	Representation
Chroma	Mean + SD^2
Root Mean Square	Mean + SD^2
Spectral Centroid	Mean + SD^2
Spectral Bandwidth	Mean + SD^2
Spectral Rolloff	Mean + SD^2
Zero Crossing Rate	Mean + SD^2
Mel Frequency Cepstral Coefficients	Mean + SD^2
Harmony	Mean + SD^2
Tempo	Mean

Table 2: Feature Representation

tion. if so, a one-way ANOVA accompanied by a pairwise Tukey HSD will look into all comparisons involving k-NN, and show which model is more significantly accurate. For the second hypothesis, a qualitativer approach will count the amount of classifiers that confuse a pair of subgenres is at least two.

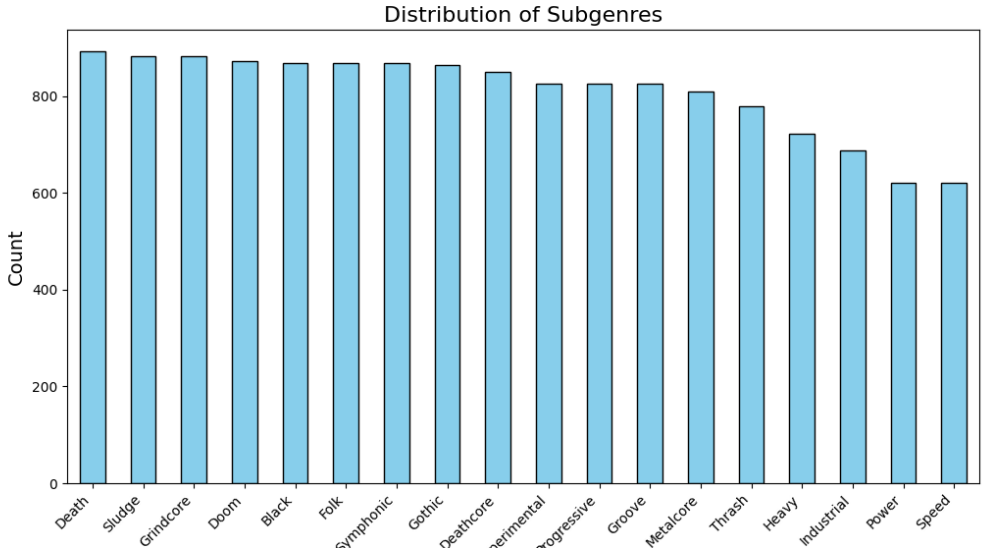

Figure 1: Subgenres are not evenly distributed due unavailability in the UK

5 Results

The SVM classifier did not finish running and would ignore `SIGINT` (signal interrupts), hence, it was omitted from the results.

5.1 First Hypothesis

The P-values of the Shapiro-Wilk Test in Table 3 suggest that the accuracy values are normally distributed (P-value > 0.05). Hence, a one-way ANOVA test is done, which has P-value of $3.449\,88 \times 10^{-166}$. This shows that there is a significant difference between the models' accuracies.

Lastly, using the results from Tukey HSD, which shows that every pairwise comparison is significant, it can be found that k-NN is only significantly more accurate than MLP. Consequently, the first null hypothesis cannot be rejected.

Model	Accuracy Mean (%)	Accuracy Std	F1 Score Mean (%)	F1 Score Std	P-value
Logistic Regression	12.6	0.006592	8.2	0.004979	0.28361
Random Forest	18.9	0.006340	10.7	0.005235	0.13833
k-NN	10.5	0.004985	11.2	0.005436	0.26110
Naïve Bayes	14.0	0.006732	10.6	0.006701	0.62132
MLP	4.3	0.002944	0.8	0.001565	0.49957
Deep	15.7	0.008363	12.5	0.008596	0.11413

Table 3: Performance Metrics of Models and P-values of Shapiro-Wilk Test

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
Deep	Logistic Regression	-0.0308	0.0	-0.0351	-0.0265	True
Deep	MLP	-0.1135	0.0	-0.1178	-0.1092	True
Deep	Naïve Bayes	-0.0172	0.0	-0.0215	-0.0129	True
Deep	Random Forest	0.0326	0.0	0.0283	0.0369	True
Deep	k-NN	-0.052	0.0	-0.0563	-0.0477	True
Logistic Regression	MLP	-0.0826	0.0	-0.0876	-0.0777	True
Logistic Regression	Naïve Bayes	0.0136	0.0	0.0087	0.0186	True
Logistic Regression	Random Forest	0.0635	0.0	0.0585	0.0684	True
Logistic Regression	k-NN	-0.0212	0.0	-0.0262	-0.0163	True
MLP	Naïve Bayes	0.0963	0.0	0.0913	0.1012	True
MLP	Random Forest	0.1461	0.0	0.1411	0.151	True
MLP	k-NN	0.0614	0.0	0.0565	0.0664	True
Naïve Bayes	Random Forest	0.0498	0.0	0.0449	0.0548	True
Naïve Bayes	k-NN	-0.0348	0.0	-0.0398	-0.0299	True
Random Forest	k-NN	-0.0847	0.0	-0.0896	-0.0797	True

Table 4: Multiple Comparison of Means - Tukey HSD, FWER=0.05

5.2 Second Hypothesis

The confusion matrices are available in the appendix from Figure 2. The matrix of k-NN and Random Forest both show asymmetric confusions with the following percentages respectively:

- Symphonic and Folk - 75%, 46%
- Power and Speed - 56%, 50%

Therefore, there are two pairs of subgenres with more than 40% confusion in at least two classifiers. Through qualitative observation, the second null hypothesis is rejected.

6 Discussion

The results provide interesting insights. There is an expected drop-off that shows that the accuracies of automatic classification of subgenres is lower compared to genres. This was drop-off expected due to the subgenres having similar traits (by extension the features in the dataset) to a parent genre [7], but the in this case the difference is high (from 55-82%). These results are consistent with those reported by Tsatsishvili et al. [17], further reinforcing that classification of the metal subgenre demands more nuanced feature extraction and analysis. Additionally, the order of the classifiers ranked by mean accuracies found by Ndou et al. [13] is k-Nearest Neighbours, Multilayer Perceptron, Random Forest, Support Vector Machine, CNN, and Logistic Regression while this experiment found a completely unexpected permutation of Random Forest, CNN, Naive Bayes, Logistic Regression, k-NN, and MLP. Finally, the mean accuracies and the confusion matrix of the MLP model suggest potential issues in the training or cross-validation process, indicating that these steps may not have been conducted optimally.

Looking at the musicological properties of the subgenres can hint as to why they were asymmetrically misclassified. Speed Metal emerged in the early 1980s. It combined the stylistic elements of the New Wave of British Heavy Metal (NWOBHM) with the raw intensity of hardcore punk. The latter, popularized by bands like Black Flag in the early 1980s, emphasized rhythm-focused songwriting, fast tempos, shouted vocals, and a characteristic drum pattern known as D-beat. Speed metal laid the groundwork for the development of two distinct genres: the heavier and more aggressive Thrash Metal, and Power Metal. Specifically, Power Metal borrowed rhythmical structure, fast tempo, and extensive usage of two bass drums from Speed Metal [17]. That is quite of lot of similar traits, so it is no surprise that these subgenres where asymmetrically misclassified.

Folk Metal and Symphonic Metal share a common thread in their ability to merge Metal with other musical traditions, creating layered and atmospheric compositions with themes of mythology and fantasy. Folk metal employs instruments like violins and bagpipes, alongside traditional melodies rooted in cultural histories. Symphonic metal uses orchestral instrumentation, including strings, choirs, and keyboards, to create a sweeping, dramatic soundscape [10]. Due to the likely instrumental and thematic overlap, these two subgenres were also asymmetrically misclassified.

7 Conclusion & Future Work

While this experiment focuses on metal subgenres, future work should explore other musical subgenres to assess if they can be more reliably classified with this methodology. Moreover, unlike the results from [13], the accuracies could potentially improve by using longer temporal slices like 30 seconds and comparing advanced tree-based ensemble learning methods. Specifically, Decision Trees, Random Forests, Extremely Randomized Trees as seen in [1]. Lastly, the dataset in this study could have had the same track labeled as multiple subgenres. Instead, aggregating the perspectives of several annotators could reduce bias and provide a more accurate representation of subjective characteristics. Leveraging inter-annotator agreement measures will also ensure consistency and objectivity in the

dataset.

To conclude, the experiment's outcomes underscore the complexities inherent in subgenre classification, driven by nuanced musical overlaps and subjective genre definitions. This study explored the classification of metal subgenres using machine learning and deep learning methodologies, employing a custom dataset curated from Spotify's public playlists. While the results showcased some expected challenges, the findings also highlighted the limitations of existing non tree-based ensemble classification techniques, indicating that k-NN is not the most accurate model for subgenre classification, as Random Forest demonstrated highest accuracy. The second hypothesis was validated, revealing significant confusion between certain subgenre pairs across classifiers, such as Folk Metal and Symphonic Metal, and Speed Metal and Power Metal, due to their shared musicological properties.

To reflect, this research reaffirmed the importance of both domain knowledge and musicological understanding in solving real-world problems. It has deepened my appreciation for the interdisciplinary nature of this field and left me eager to explore tree based ensembles to further improve the accuracy and to try to properly annotate a music dataset with knowledgeable specialists in the field of musicology.

A Appendix

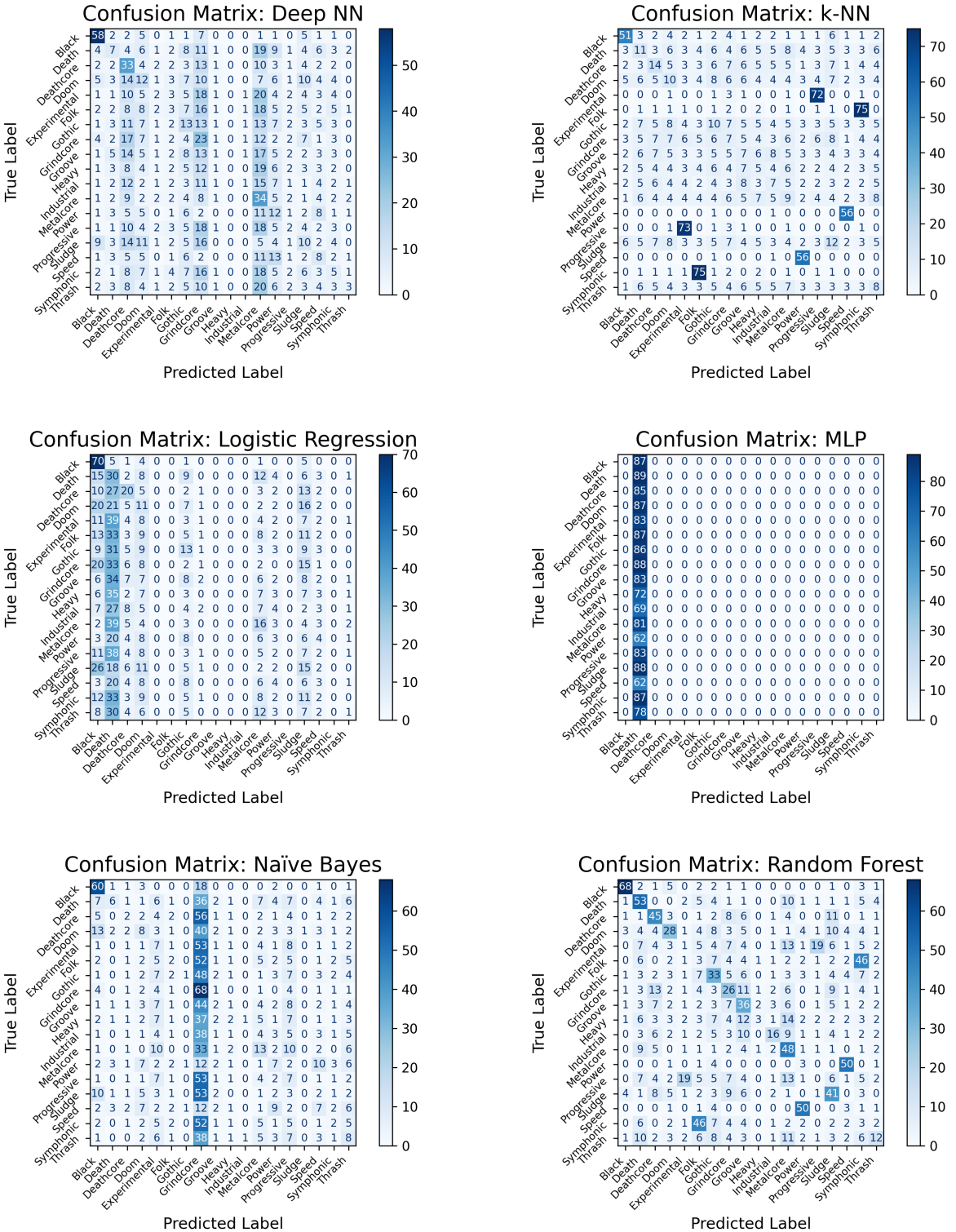

Figure 2: Confusion matrix of each classifier

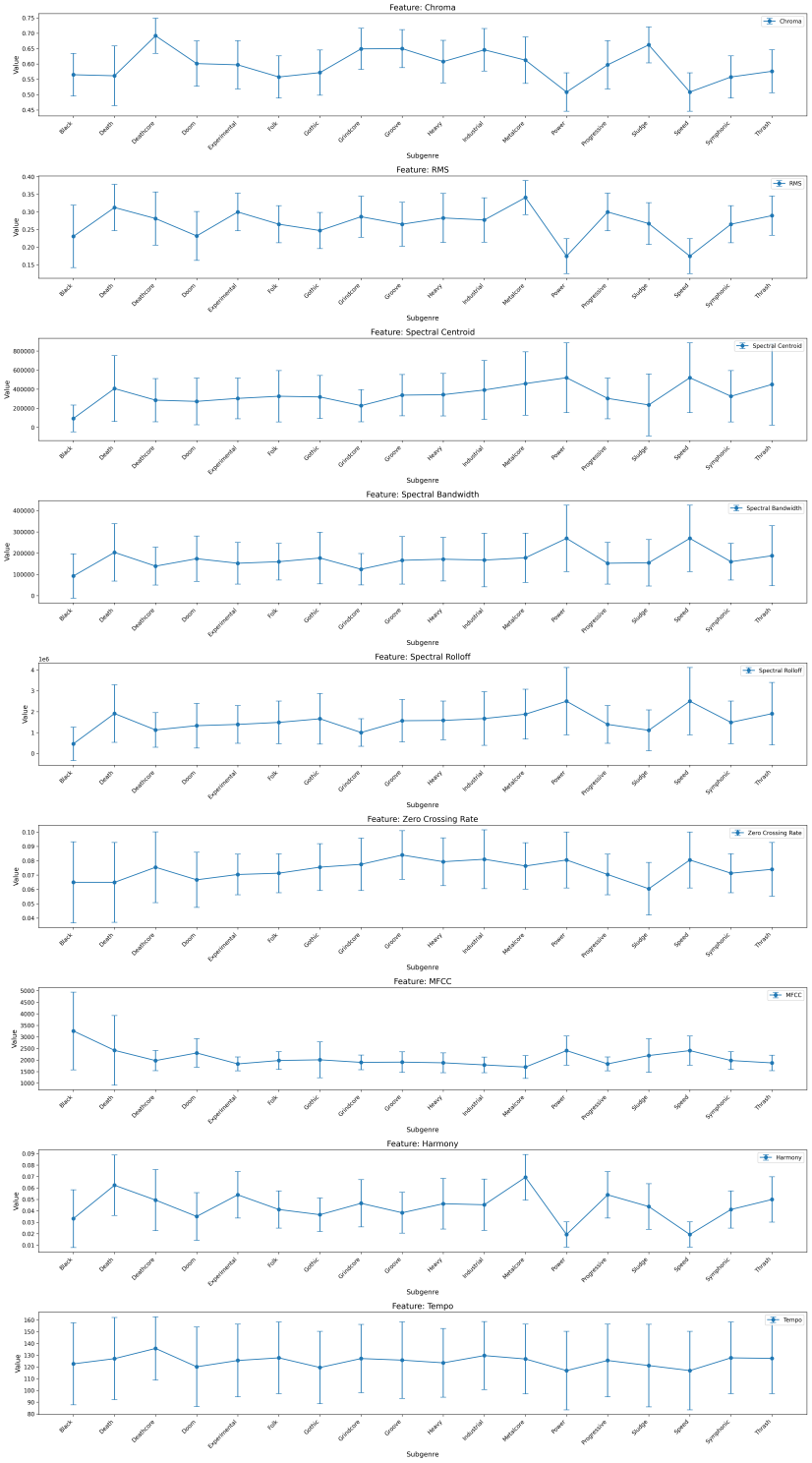

Figure 3: Feature differences per subgenre

References

[1] Laura Pérez-Molina Antonio Caparrini, Javier Arroyo and Jaime Sánchez-Hernández. Automatic subgenre classification in an electronic dance music taxonomy. *Journal of New Music Research*, 49(3):269–284, 2020.

[2] Sonia Archer-Capuzzo and Guy Capuzzo. *Metaldata: A Bibliography of Heavy Metal Resources*, volume 43. AR Editions, Inc., 2021.

[3] Mariana Lopes Barata and Pedro Simoes Coelho. Music streaming services: understanding the drivers of customer purchase and intention to recommend. *Heliyon*, 7(8), 2021.

[4] Roger B Dannenberg. Style in music. *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, pages 45–57, 2010.

[5] Maria Eriksson, Rasmus Fleischer, Anna Johansson, Pelle Snickars, and Patrick Vonderau. *Spotify teardown: Inside the black box of streaming music*. Mit Press, 2019.

[6] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[7] Philip Hider and Deborah Lee. A polyphony of characteristics: An analysis of the categorisation of music's subgenres. *Journal of Information Science*, page 01655515231203511, 2023.

[8] Tim Ingham. Over 60,000 tracks are now uploaded to spotify every day. that's nearly one per second. *Music Business Worldwide*, 24, 2021.

[9] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE, 2013.

[10] Peter A Marjenin. *The metal folk: The impact of music and culture on folk metal and the music of Korpiklaani*. Kent State University, 2014.

[11] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, pages 18–24, 2015.

[12] Kembrew McLeod. Genres, subgenres, sub-subgenres and more: Musical and social differentiation within electronic/dance music communities. *Journal of popular music studies*, 13(1):59–75, 2001.

[13] Ndiatenda Ndou, Ritesh Ajoodha, and Ashwini Jadhav. Music genre classification: A review of deep-learning and traditional machine-learning approaches. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–6. IEEE, 2021.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[15] Nikki Pelchat and Craig M Gelowitz. Neural network music genre classification. *Canadian Journal of Electrical and Computer Engineering*, 43(3):170–173, 2020.

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

[17] Valeri Tsatsishvili. Automatic subgenre classification of heavy metal music. Master's thesis, 2011.

[18] Deena Weinstein. *Heavy metal: The music and its culture*. Da Capo Press, 2000.