Evolution of Indie EDM

Jacob Bengel
Kansas State University, Manhattan, KS 66502
jbengel@ksu.edu

December 11, 2024

Abstract

This paper investigates the evolution of electronic dance music (EDM), especially smaller/Indie artists, during recent years. An original dataset of 2,060 songs was curated from the Youtube channels of 27 musicians while adhering to specific criteria to provide basic standardization and reduce the likelihood of noise and egregious outliers being included. The open-source Essentia toolset was used to extract over 2,000 features which were narrowed down using Scikit to allow for supervised machine learning analysis using Weka. Random Forest and Logistic regression were successfully used to classify songs within the dataset to the appropriate artists far better than random chance. The genre of EDM was observed to have evolved across the span of 2010-2024 while providing for the continuation of various subgenres. The challenges entailed by sourcing audio data from Youtube are explored and it is thus advised that future work intending to be more thorough should consider making use of pre-existing & actively-curated music datasets.

1 Introduction

The origins of what came to be recognized as "electronic dance music" (EDM) trace back to the 1960s, spanning Disco, Synth, and various burgeoning electronic techniques & styles. The 1990s broadened this soundscape to include "techno music, house music, [hardstyle], dub, trance, and drum and bass" (The Los Angeles Film School, 2017). The 2000s furthermore introduced techno-pop, dubstep, house, trap, and yet more genres as artists remixed and sampled a virtually-endlessly increasing variety of sounds. The EDM landscape today features faces (and non-faces) such as Avicii, Martin Garrix,

Daft Punk, Marshmello, Deadmau5, Skrillex, and hundreds more (Ranker, 2024). The increased interest and production of EDM has led to contention over what it is that specifically defines the genre whilst others may scoff and dismiss it as an innovative form of noise pollution.

This analysis endeavors to explore trends in EDM across the previous decade through the lens of data science and so it is prudent to briefly cover previous, related work. Van der Velden and Hitters (2016) performed a qualitative analysis of EDM as a genre, finding it to be incompatible with established musicological models and thus warranting the development of a new analytical framework. Difficulties in analyzing EDM compared to more mature or more conventional genres isn't unexpected given the broad scope and individualistic nature of EDM as a genre. Bogdanov et al. (2013) developed Essentia, an open-source toolset for music and audio analysis which has come to be widely used in research on music analysis and music information retrieval (MIR). Essentia is still actively maintained and was *essential* to the generation of the dataset used in this paper. Knees et al. (2015) would go on to employ Essentia in the generation of two novel datasets to be employed for EDM analysis. Faraldo et al. (2016) used Essentia in an involved effort to classify the tonality of a set of EDM samples. Another paper by Eigenfeldt and Pasquier (2013) found success in generating EDM tracks through machine learning techniques.

The realities of compiling and analyzing an EDM dataset entail (non-exhaustively) a number of challenges: identifying a source (or multiple), interfacing with said source, reasoning about what datapoints to include/exclude, and the manpower requirements of vetting enough entries to make for a useful machine learning analysis. This project originally intended to source data from Spotify, which might have simplified collecting songs according to more nuanced parameters (e.g. all songs of the top artists in a genre over a timespan), but, due to difficulties using the API, opted to instead use Youtube as a source, which required working through a different set of limitations. This necessarily narrowed the scope to primarily consist of "Indie EDM" due to the unavailability of data from the channels of mainstream artists that *might* better represent "EDM" as a whole. This is, however, a niche the author has more familiarity with, though not subject matter expertise. This paper hypothesizes that the (Indie) EDM genre retains a consistent signature throughout the years since 2010, but that a select few have drastically shifted their style.

2 Methods

In accordance with the resource constraints of this project, and to provide novelty compared to existing datasets, songs were sourced from artists' channels on Youtube. An initial list of artists was drafted including those familiar to the author or simply popular within the genre; additional artists were identified for inclusion in this list by referencing collaborations, competitions, remixes, or declared inspirations. Approximately half of the artists/channels on the initial list were excluded due to unclear alignment with the EDM genre or, primarily, having channel content unsuited to MP3 conversion e.g. high ratios of non-music content, uploads that are non-representative of their style, or content that otherwise complicates discerning which items are original music. Further requirements were adhered to in selecting individual songs for inclusion in the dataset:

- The artist must have at least 12 qualifying uploads in total
- The artist self-identifies with the EDM genre or would plausibly be recognized as part of it
- 90 seconds minimum length
- Uploaded no earlier than 2010
- No extraneous non-song audio e.g. as part of a music video
- No remixes of other artists' songs
- Collaborations, competition entries, and experimental tracks are okay

Version 2024.11.18 of the tool YT-DLP was used to download 2,060 songs selected according to these standards as MP3 files. As might have been alleviated by sourcing music through the Spotify API or an established/commercial source for music data, certain limitations of this dataset are acknowledged:

- The selection of artists is not random and is skewed towards smaller/Indie creators
- The identity of some artists as representative of EDM is debatable
- Uploaded Youtube content may be non-representative of artists' other works
- Non-EDM songs were not manually listened to and removed

- Audio contents of converted MP3 files may contain "watermarking" that unintuitively biases machine learning techniques and may thus over-represent classifcation capabilities
- Non-zero human error in compiling the dataset leading to erroneous inclusions & exclusions
- Non-correlation of upload dates with original release dates of old tracks, especially those by Daft Punk
- Existence of duplicate entries for collaborations under different artists

The set of MP3 files was processed using a feature extractor provided by Essentia 2.0.1 and converted into a CSV format also containing artist names, video upload IDs, upload dates, view counts, and song titles. Scikit 1.6.0 was used to perform One-Class SVM outlier detection, which eliminated approximately 200 outliers without manual effort, and to apply Principal Component Analysis to select for the 250 most informative of the more than 2,000 features identified by Essentia. Manual correction of the new "reduced" dataset eliminated 4 different duplicate entries (by upload ID) believed to be caused by a bug with YT-DLP.

The initial 2,060 MP3 files occupied 9.6 GiB of storage space, averaging 4.4 MiB per file. They would take 133.5 hours to play back-to-back, averaging 2 minutes and 53 seconds per song. The `reduced.csv` file containing the processed features from Essentia occupies just 7.6 MiB. As seen in Figure 1, of the 27 total artists, Hinkik, DJ Nate, and Xomu contribute the fewest songs to the dataset (fewer than 25 each), while Riff Kitten, Panda Eyes, and Waterflame contribute the most (more than 125 each). The average artist in this dataset uploaded $\overline{x} = 68.4$ (non-outlier) songs with a standard deviation of $\sigma = 51.9$. As seen in 2, the song count per year is a bimodal distribution with local maxima in 2015 and 2022 (note: the data for 2024 does not include December). There were an average of $\overline{x} = 123.1$ songs per year with a standard deviation of $\sigma = 52.8$.

Initial exploration of the dataset was performed using Weka 3.8.6. A new CSV file had to be created to 1) reduce the number of columns to 100 for performance reasons, prioritizing the most informative features, and 2) remove filenames as the presence of certain characters prevented Weka from fully reading the file. All using 10-fold cross-validation, `ZeroR` classification correctly classified the **artist** of 15.5% of the instances; `OneR` correctly classified 17.0% of instances; `RandomForest` correctly classified 50.9% of instances; and `SimpleLogistic` correctly classified 59.2% of the total instances. Refer to the confusion matrix in Table 1. 59.2% accuracy using `SimpleLogistic`

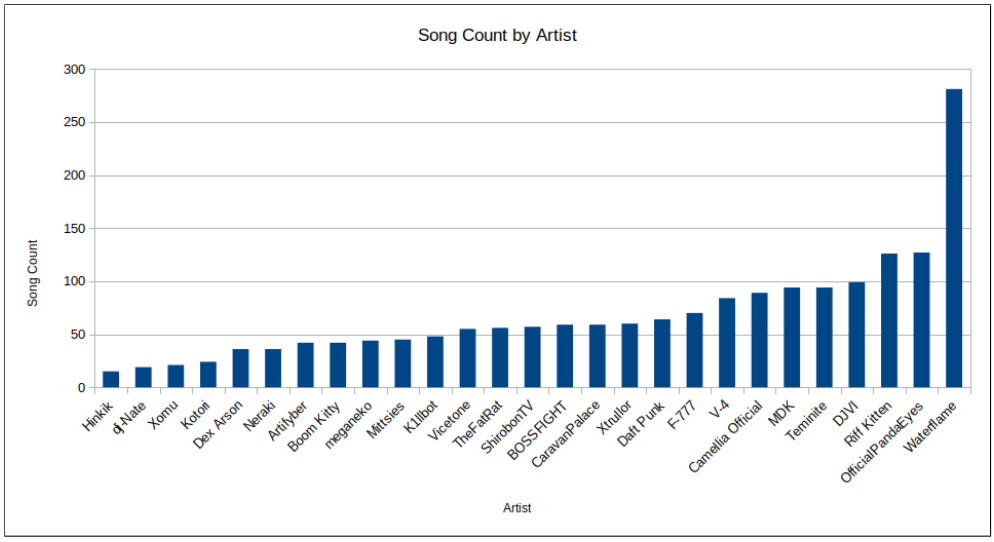

Figure 1: Bar chart displaying number of songs uploaded by each artist.

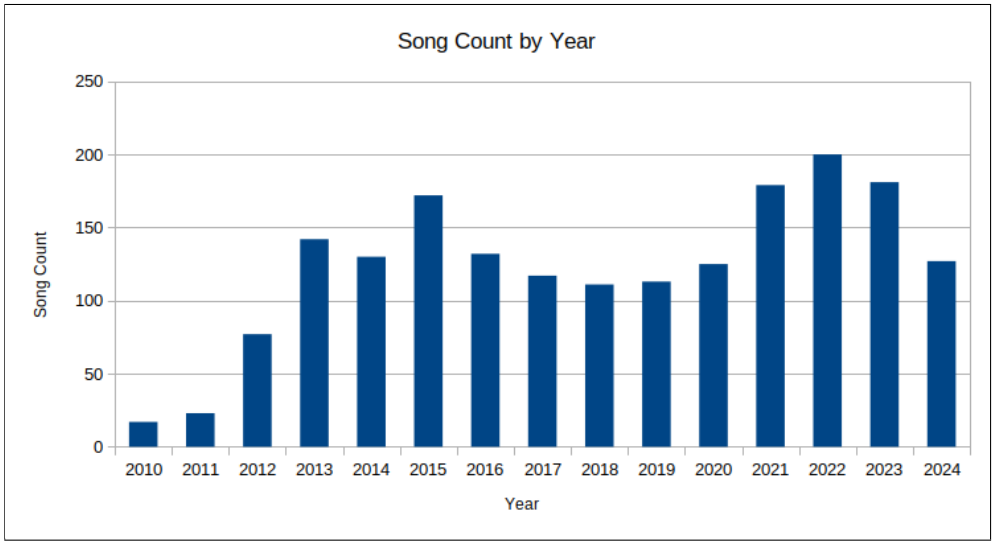

Figure 2: Bar chart displaying number of songs uploaded each year from 2010 to 2024.

compared to the 15.5% of randomly guessing using `ZeroR` is promising and strongly suggests the existence of meaningful patterns within the dataset to distinguish between artists' styles. Also all using 10-fold cross-validation, `ZeroR` classification correctly classified the **year** of 10.8% of the instances; `OneR` correctly classified 11.9% of instances; `RandomForest` correctly classified 24.5% of instances; and `SimpleLogistic` correctly classified 20.4% of the total instances. In this case, `RandomForest` outperformed `SimpleLogistic`, while both performed much better than random, this result is likely skewed by disproportionate contributions to the dataset by artists across the years. Refer to the confusion matrix in Table 2.

Artifyber	**17**	0	0	4	1	1	0	1	1	3	0	0	0	0	1	2	2	3	1	0	0	0	1	0	3	0	1
BoomKitty	0	**24**	1	1	0	1	0	0	0	1	0	0	0	1	1	0	2	1	0	3	5	0	0	0	1	0	0
BOSSFIGHT	1	2	**19**	3	0	1	1	1	3	0	1	0	1	6	2	1	1	3	0	3	1	0	0	1	3	0	5
CamelliaOfficial	3	1	1	**74**	0	0	1	0	0	0	0	0	0	1	1	0	1	1	0	2	2	0	0	0	0	1	0
CaravanPalace	1	0	0	1	**38**	3	0	0	0	0	0	0	1	0	0	1	0	1	4	0	1	3	0	3	1	0	1
DaftPunk	2	0	0	0	7	**29**	0	0	2	2	0	2	0	4	0	4	0	0	6	0	0	0	1	0	2	2	1
DexArson	0	0	0	0	0	0	**15**	2	1	2	0	2	1	3	0	0	0	6	0	0	3	0	0	0	0	0	1
djNate	0	0	1	0	0	1	3	**7**	0	0	0	0	0	1	0	0	0	2	0	0	0	1	0	0	2	0	1
DJVI	1	1	0	0	0	1	1	0	**71**	3	0	0	0	0	1	0	0	1	0	1	1	2	3	3	5	4	0
F777	2	1	1	2	0	0	2	0	3	**36**	0	1	0	8	0	0	1	2	0	0	0	1	3	0	4	1	2
Hinkik	0	0	0	1	0	0	0	0	0	1	**2**	1	0	1	0	0	0	1	0	0	3	0	0	4	0	0	1
K1llbot	1	0	1	1	0	3	0	0	0	0	1	**25**	1	1	1	0	2	3	1	1	2	0	1	0	1	0	2
Kotori	0	0	3	0	0	0	0	0	0	0	0	0	**14**	0	0	0	1	5	0	0	0	0	0	0	0	0	1
MDK	0	0	2	2	1	0	1	1	4	4	1	0	1	**42**	3	1	3	4	0	2	6	1	2	0	8	0	5
meganeko	0	0	3	1	0	0	0	0	2	0	1	0	0	3	**12**	1	1	4	0	3	2	4	0	1	5	0	1
Mittsies	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	**35**	0	0	2	1	0	0	1	0	3	0	0
Neraki	0	2	0	3	0	0	0	0	0	3	0	0	2	2	0	0	**14**	0	0	0	5	0	0	0	4	0	1
OfficialPandaEyes	0	0	4	1	3	2	6	1	0	6	0	1	3	6	1	1	1	**73**	1	0	6	5	0	2	1	0	3
RiffKitten	0	0	0	0	6	9	0	0	0	0	0	0	0	0	0	0	0	2	**83**	2	2	2	2	4	14	0	0
ShirobonTV	3	0	1	2	3	1	0	1	1	0	0	0	1	4	3	1	0	3	0	**23**	0	3	0	0	7	0	0
Teminite	0	2	2	2	0	0	5	0	5	1	0	0	1	4	0	0	2	9	3	0	**51**	2	0	1	2	1	1
TheFatRat	0	0	0	0	4	1	0	1	2	0	0	0	0	4	2	0	0	5	6	2	0	**19**	0	6	2	2	0
V4	0	3	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	0	0	1	**68**	0	7	0	1
Vicetone	0	0	0	0	2	0	0	0	2	0	1	1	0	0	1	0	0	1	3	1	0	4	0	**36**	0	2	1
Waterflame	1	3	2	4	1	4	0	1	4	1	0	0	0	1	1	4	2	0	14	2	2	1	6	0	**231**	0	1
Xomu	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	3	1	0	2	4	0	3	0	**6**	0
Xtrullor	0	0	0	1	0	2	1	0	0	0	0	1	1	3	3	0	4	3	0	0	3	0	1	2	2	1	**32**

Table 1: Confusion matrix for artists using 10-fold cross-validation SimpleLogistic classification.

2010	**0**	0	2	2	3	4	0	0	0	0	1	3	1	0	1
2011	0	**0**	1	4	3	9	1	0	0	0	0	1	3	1	0
2012	0	1	**9**	10	2	15	5	1	1	2	5	7	13	6	0
2013	0	0	5	**51**	12	33	7	2	1	0	2	5	16	8	0
2014	0	0	3	24	**22**	30	2	7	1	6	6	8	14	6	1
2015	0	0	1	15	15	**63**	15	6	0	11	7	14	14	5	6
2016	0	0	2	10	4	31	**23**	7	6	6	4	15	16	5	3
2017	0	0	0	10	6	19	16	**11**	3	4	11	13	15	8	1
2018	0	0	0	8	8	16	9	2	**13**	8	5	12	17	11	2
2019	0	0	1	2	3	14	8	4	4	**19**	10	23	16	5	4
2020	0	0	2	4	3	11	7	2	1	11	**34**	22	14	12	2
2021	0	0	1	10	7	12	8	2	5	9	13	**49**	30	29	4
2022	0	0	0	10	5	15	10	5	5	4	11	24	**86**	22	3
2023	0	0	3	15	10	9	4	3	3	3	11	34	27	**54**	5
2024	0	0	0	4	4	14	5	0	2	6	6	16	25	27	**18**

Table 2: Confusion matrix for years spanning 2010-2024 using 10-fold cross-validation RandomForest classification.

Although Scikit has already been used to select for informative features used in the dataset, using Weka to perform further feature selection using `Correlation`, `InfoGain`, and `OneR` attribute evaluation only commonly include `lowlevel.erbbands.median.4` among the 10 most informative features.

3 Results

From this low-cost and high-manual-effort dataset, it was found that, assuming the analysis was not derailed by a poor selection of a data source, that the represented artists styles' are quantifiably distinct from one another. Less compelling, but still better than random, evidence suggests that the broader EDM genre has evolved in the years since 2010 (though certainly since the 1960s).

Using Weka to perform `SimpleKMeans` clustering up to 7 clusters unilaterally resulted in incorrect clustering in at least 80% of instances across all the cluster counts. The 2-cluster clustered around Riff Kitten and Waterflame which marks a distinction between electro-swing and chiptune/video-gamey music. N-clusters 3 through 5 additionally clustered around DJVI, MDK, and Panda Eyes, respectively, introducing more "typical" loud & aggressive EDM styles. It was unexpected that it would take until 7 clusters to separate

out Daft Punk as its own cluster given that are presumed to be the most mainstream artist within the dataset and their style is much more laid-back compared to the other artists.

A specific observation that inspired this choice of paper was that Bossfight's older music is drastically different than what they produce today (far less happy-go-lucky/energetic and much more loud/dramatic) - and this may not be a novel observation - but classification using Weka validates this line of thinking. As Bossfight has only 59 entries in the dataset (3.9 per year) and there is a rather telling gap between 2012 and 2015, their entries were manually clustered to a 2011-2012 class, a 2015-2019 class, and a 2020-2024 class containing their newest tracks. `ZeroR` classification correctly classified 40.7% of instances while `SimpleLogistic` classified a far greater 86.4% to the correct cluster.

4 Conclusion

The two primary findings of this paper are that (Indie) EDM artists' styles are not merely qualitatively distinct from one another, and that the artist Bossfight has indeed redefined their musical signature. Features selected for within the reduced dataset may (or may not) point toward a greater understanding of EDM as a genre for those knowledgeable of the musical/audio meaning of the features extracted by Essentia.

The Achilles' heel of this analysis is whether or not sourcing MP3 files from Youtube, or by using YT-DLP specifically, introduced unintended patterns such as an audio watermark that oversimplified the classification process; it is uncertain how introducing data from a different source such as Spotify would affect this. A possible solution for this would be to pay for access to a professionally-maintained database, which would also potentially be better suited to random sampling rather than arbitrary identification & selection of artists & songs. Essentia's website links to applications of their tooling such as AcousticBrainz, which crowdsources acoustic data; making use of such already-available data might also be a major improvement upon the methodology established in this paper. Expansion upon the topic of EDM in particular could incorporate a broader variety of artists, though it is doubtful that continuing to use Youtube as a source for MP3 files would be prudent.

Acknowledgments

Dr. Lior Shamir for a job well done teaching the concepts of Data Science Foundations during the Fall 2024 semester at Kansas State University and providing references on Essentia and LaTeX use.

References

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). Essentia: an open-source library for sound and music analysis. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 855–858.

Eigenfeldt, A. and Pasquier, P. (2013). Evolving structures for electronic dance music. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, pages 319–326.

Faraldo, Á., Gómez, E., Jordà, S., and Herrera, P. (2016). Key estimation in electronic dance music. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 335–347. Springer.

Knees, P., Faraldo Pérez, Á., Boyer, H., Vogl, R., Böck, S., Hörschläger, F., Le Goff, M., et al. (2015). Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR); 2015 Oct 26-30; Málaga, Spain.[Málaga]: International Society for Music Information Retrieval, 2015. p. 364-70.* International Society for Music Information Retrieval (ISMIR).

Ranker (2024). The greatest edm artists of all time.

The Los Angeles Film School (2017). A brief history of edm. *The Los Angeles Film School*.

Van der Velden, J. and Hitters, E. (2016). The distinctiveness of electronic dance music. challenging mainstream routines and structures in the music industries. *International Journal of Music Business Research (online)*, 5(1):59–84.