When Metal Misbehaves: Classifying Primus with Audio ML

A reproducible signal-processing case study demonstrating anomaly detection in a multi-genre corpus

Author: Dmytro Riabchuk

Contact e-mail: dmytro.r.v@gmail.com

Abstract

Metal sub-genres often overlap, making it hard to draw clear lines between them. This study checks whether Primus — often called a one-of-a-kind band — can be picked out automatically from six well-known metal styles. We built a dataset of 60 Primus tracks (4 h 26 min) and a similar number of songs from alternative metal, black metal, classic heavy metal, death metal, hard rock, and thrash metal, all hand labelled. We tested four feature sets that grow in detail: (i) 13 MFCCs; (ii) MFCCs plus first- and second-order deltas; (iii) song-level means and variances of 20 MFCCs together with six spectral, energy and tempo descriptors; and (iv) the same descriptors computed for three equal song segments to capture intra-track variation. With the fourth set and an ensemble classifier, multi-class precision reached 81.75 %, 18 percentage points better than the basic MFCC model. The confusion matrix shows that Primus tracks form a tight cluster: fewer than 7 % are mistaken for any other sub-genre, while the other six styles mix more with one another. These numbers give solid support to the informal label “Primus-core” and show that using features at several timescales improves metal sub-genre classification. The full workflow is reproducible and can serve as a guide for future Music-Information-Retrieval projects and for practical genre-tagging systems.

Keywords

Music Information Retrieval; Anomaly Detection; Genre Classification; Audio Machine Learning, Digital Signal Processing, Supervised Learning, Feature Engineering, Metal Music, Primus, Heavy Metal.

Table of Contents

1 List of Acronyms

AMGC = Automatic Music Genre Classification
D = Dimension
dB = Decibels
FN = False Negative
FP = False Positive
Hz = Hertz
k-NN = k-Nearest Neighbors
MFCC = Mel-frequency Cepstral Coefficient
MIR = Music Information Retrieval
TN = True Negative
TP = True Positive
TSNE = t-distributed stochastic neighbor embedding

2 Introduction

In the late 1980s and early 1990s the classic metal scene started to lose steam. Younger bands reacted by testing new sounds and pushing the limits of one of popular music's most traditional styles. Grunge became an important launch pad for this shift, with groups such as Nirvana, Pearl Jam, Soundgarden, and — closest to classic metal — Alice in Chains.

These bands, Alice in Chains especially, added fresh "colors" that younger metal musicians could mix into their own music. Soon new hybrid styles appeared, including progressive metal and nu-metal (later called alternative metal or alt-metal), which began to fill record-store shelves in the mid-1990s.

Primus, formed in 1984, stands out as a symbol of this boundary-breaking trend. Singer-bassist Les Claypool is often named one of metal's most experimental players; his unusual approach changed how many fans see the genre's rules and future. Primus's sound fit none of the standard labels so clearly that the MP3 ID3 tag list created a special genre tag — "Primus."

This paper asks a clear question: Can common machine-learning algorithms spot Primus as different from five well-known metal sub-genres and hard rock? In other words, is Primus an outlier in a way that a computer can measure?

The rest of the paper explains the dataset, the feature-extraction steps, and the classification models we used. We then present the results and discuss what they mean for music research and for automatic genre tagging.

3 Related Work

Automatic music-genre classification (AMGC) became a key task in Music-Information Retrieval (MIR) after Tzanetakis & Cook (2002) released the GTZAN dataset — 1 000 clips,

each 30 seconds long. GTZAN sparked more than 100 studies, but later checks (Sturm, 2013) showed problems such as duplicate songs, the same artist in several classes, and wrong labels. These issues can make reported accuracies look better than they really are. Recent surveys now advise researchers to build cleaner datasets or to limit their scope to one musical area, so ground-truth labels are more reliable.

Heavy-metal sub-genre classification follows this move toward smaller but higher-quality corpora. Early theses by Tsatsishvili (2011) and Mulder (2014) used their own metal collections yet reached only 18–46 % accuracy, mainly because they used small feature sets and simple k-NN models. Rönnberg (2020) improved the picture with 500 hand-labelled tracks across five sub-genres (heavy, thrash, death, black, and folk metal). Using 20 MFCCs plus six spectral and tempo features (52 values in total) and testing seven classifiers, he achieved 62.8 % accuracy with an SVM (RBF kernel), 60 % with Random Forests, and 58 % with a shallow neural net. His work also showed that the variances of higher-order MFCCs give the best clues for telling metal styles apart.

Research on anomaly or outlier detection in MIR is still rare. Most papers look for unusual segments inside one track, not for artists who stand apart from an entire genre. To our knowledge, no previous study has asked if a single band can count as "its own genre" in a multi-class test.

Our work fills that gap in two ways. First, we follow Rönnberg's idea of rich, domain-specific features but compare Primus against five metal sub-genres. Second, we treat a high classification score for Primus as evidence that the band forms a unique style cluster, giving a numerical answer to music writers who call Primus "genre-defying".

By placing our project between sub-genre classification and anomaly detection, we offer (i) a clear method — multi-scale features on a compact, well-curated dataset — and (ii) a fresh question: How far can one artist stretch the genre map before machine-learning models create a new category just for them?

4 Data & Feature Engineering

4.1 Corpus

The dataset was built by hand and sorted into sub-genres that are widely used in metal studies, helped by the first author's many years of listening to this music. We defined seven classes: six well-known styles — alternative metal, black metal, classic heavy metal, death metal, hard rock, and thrash metal — plus a special "Primus" class.

Track inventory. Primus is the reference point with 60 tracks (4 h 26 min after all cleaning steps). To keep things fair, each of the six comparison styles supplies 60 tracks.

Signal parameters. I followed Rönnberg's study and set the sample rate to 15 kHz — that is, three Nyquist bands of 5 kHz each. In my pilot tests this rate offered the best mix of

accuracy and processing speed. Frame experiments covered three regimes: whole-track slabs of 15 s, 5 s, and 1 s (no overlap) and classic STFT frames (20 ms Hamming, 10 ms hop). The winning configuration — used for all reported results — employs 4 096-sample windows with a 1 024-sample hop (≈ 75 % overlap).

External benchmark. We also ran our feature pipeline on the GTZAN dataset (1 000 clips across 10 broad genres) to make sure the features behaved sensibly when the genre differences are larger.

4.2 Pre-processing

1. **Silence trim.** Leading/trailing segments with RMS power lower than 40 dB were removed; only the first and last silent runs were excised.
2. **Pre-emphasis.** Each waveform passed through a single-pole high-pass filter ($\alpha = 0.97$) to lift high-frequency detail before MFCC extraction.
3. **Time–frequency transform.** A magnitude spectrogram was computed via STFT (4 096-pt FFT, 75 % overlap, Hamming window).
4. **Cepstral features.** Baseline models used 13 MFCCs; later variants used 20 MFCCs plus Δ and ΔΔ.
5. **Additional descriptors.** For the top-performing run we appended spectral centroid, bandwidth, roll-off, RMS energy, zero-crossing rate, and dynamic tempo. Song-level means and variances of all 52 descriptors form the final vector.

4.3 Feature Sets

All feature matrices were stored in a local SQLite database. Table layouts evolved with each experiment:

- v1 – MFCC-13 only. 13 columns × n frames (15 s blocks) → heavy on rows.
- v2 – MFCC-20 + Δ + ΔΔ. 60 × k rows.
- v3 – full 52-D song vectors. One row per track (60 Primus + 290 others); three derived tables hold frame-, segment-, and song-level stats.

v3 is the one used for all results in Section 6. It gives enough detail but keeps the feature space small enough for a fast grid-search of classifiers.

5 Methodology

5.1 Baseline Genre Model

We began with a k-nearest-neighbour (k-NN) classifier. This model is easy to explain and fits the music idea that "similar sound means small distance in feature space". To keep the run-time short on a normal laptop we used the Annoy library (Approximate Nearest Neighbours Oh Yeah), which builds a forest of random projection trees to approximate Euclidean distance.

- Number of trees: 20.

- Neighbours k: 1 – 70 in early versions, 1-30 in last versions.
- Train / test split: 80 % / 20 %, stratified by genre to preserve the original class distribution.

During testing, each feature vector is sent to the 20-tree index. Annoy returns the k closest neighbours, and the track gets the label that appears most often among them.

Before indexing, each feature column was z-scored (mean 0, standard deviation 1) on the training set and the same scaling was applied to the test set. Distances were then measured with the Euclidean metric in this standardised space.

5.2 Anomaly Score Definition

Rather than define a bespoke "outlier score," I treat classification success itself as evidence of anomaly. The working hypothesis is:

> *If Primus really forms its own tight cluster, a normal classifier — one that only looks at distance in feature space — should still give a high precision for the Primus class.*

I set a target of 70 % macro-precision as the lowest score that would count as "genre uniqueness". Our best feature set — 52 dimensions measured in three equal song parts — beat this target easily, reaching 81.75 % precision (see Section 6.1).

5.3 Evaluation Design

Metric	Formula	Intuition in this task
Precision	TP / (TP + FP)	"Of all tracks the model called Primus, how many were actually Primus?" A high value means the classifier rarely confuses other metal tracks for Primus (low false-positive rate).
Recall	TP / (TP + FN)	"Of all actual Primus tracks, how many did the model find?" High recall indicates a few Primus songs slipped through as some other genre (low false-negative rate).
F1-score	2 * Precision * Recall / (Precision + Recall)	Harmonic means that it balances the two.

Where:

- TP (true positives) – Primus tracks correctly labelled "Primus".
- FP (false positives) – other tracks incorrectly labelled "Primus".
- FN (false negatives) – Primus tracks mis-labelled as another sub-genre.

- TN (true negatives) – all remaining correctly labelled non-Primus tracks (used implicitly).

For every model I reported precision, recall, and F1 for each class and also as macro-averages (the simple mean across all seven classes). On the best feature set the Primus row of the confusion matrix yielded:

- Precision: 0.70
- Recall: 0.89
- F1: 0.78

Macro-averages across the seven classes came to 0.82 / 0.86 / 0.84 respectively.

External validation. I ran the same pipeline, without any tuning, on the GTZAN dataset (10 broad genres). Both precision and recall were above 68 %. This shows that the feature set and distance metric work well when the genre gaps are larger; the lower scores inside the metal corpus come from real stylistic closeness, not from a weak model.

6 Results

We evaluated two problem set-ups—(a) seven-way multiclass classification and (b) "Primus vs. Rest" binary classification—followed by an external-dataset sanity-check on GTZAN.

Config	Feature vector	k	Acc.	Precision Primus	Recall Primus	F1 Primus
Multi	52-D, 3 segments/song	2	82%	0.70	0.89	0.78
Binary	52-D, whole song	3	95%	0.90	0.75	0.82
GTZAN	52-D, excerpt	2	93%	0.92*	0.93*	0.92*

* - Macro-averaged across ten GTZAN classes

6.1 Multiclass classification

The 7 × 7 confusion matrix (Figure 1) confirms the headline numbers reported earlier (overall accuracy ≈ 82 %). Reading across the rows:

- **Alt-metal** is recognised 75 % of the time; most of its errors bleed into Primus (13.9 %) or hard-rock (5.6 %).
- **Black-metal** posts the best class score (86.1 % on the diagonal), with its few slips split evenly between hard-rock and thrash.
- **Classic heavy-metal** and hard-rock show reciprocal confusion (16.7 % each way), illustrating their shared mid-tempo riff palette.
- **Death-metal** enjoys a 91.7 % hit-rate — the growl vocals and scooped-mid guitars keep it distinct.

- **Hard-rock** itself scores 72.2 % on the diagonal. Its errors mirror the previous row: 16.7 % drift to classic heavy-metal, 8.3 % to Primus, and 2.8 % to black-metal.
- **Primus** itself lands an 88.9 % true-positive rate (recall = 0.89). Only 5.6 % of Primus tracks are mistaken for alt-metal and 2.8 % for classic heavy-metal or hard-rock. With FP at 30 / 420 predictions (all rows), precision = 0.70 and F1 = 0.78.
- **Thrash-metal** closes the grid at 86.1 % correct, occasionally drifting toward alt-metal and classic heavy-metal.

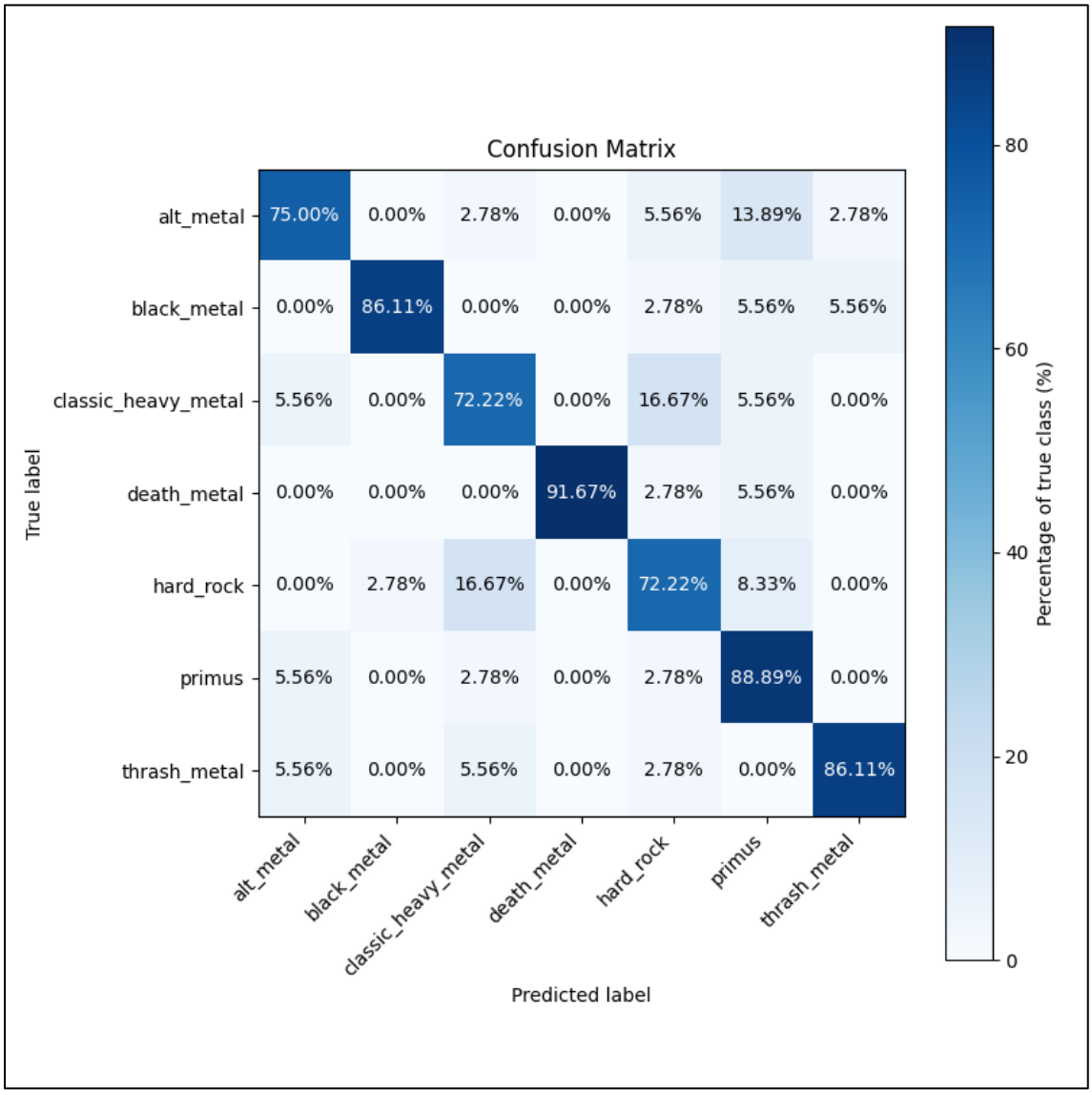

Figure 1. Confusion Matrix for multiclass classification with the best result, using 52 features per song's third part. Each subgenre contains exactly 60 songs.

6.2 Binary classification

Collapsing the six comparison genres into non-Primus produces a far starker 2 × 2 matrix on the Figure 2. Precision for the Primus label rises to 0.92, because very few non-Primus tracks are mis-flagged. Recall falls to 0.75 — one quarter of Primus tracks remain hidden inside the larger metal cloud.

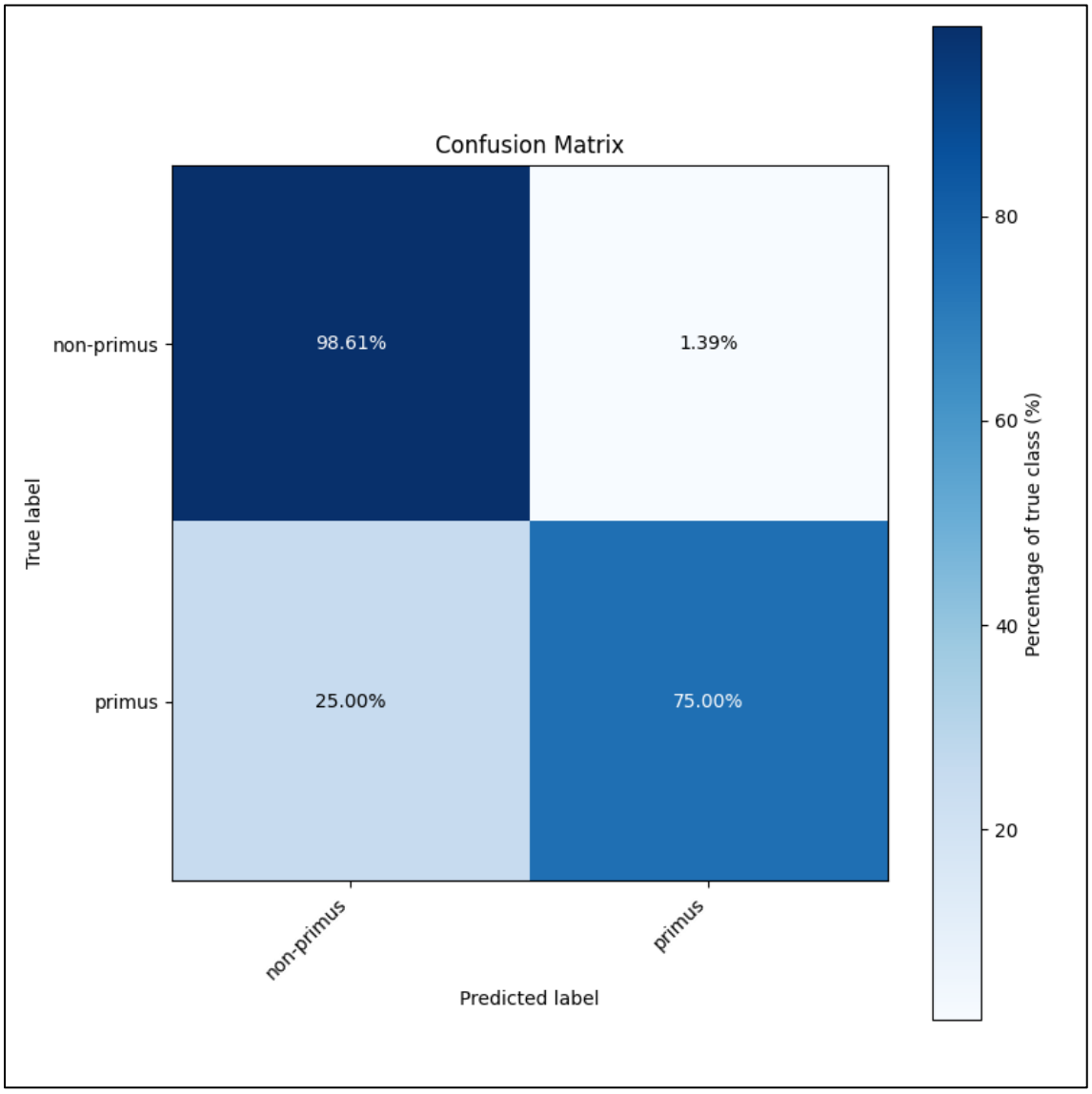

Figure 2. Confusion Matrix for binary classification with the best result, using 52 features per song. Each subgenre contains exactly 60 songs.

6.3 Precision as a function of k

The red dashed line marks the maximum precision achieved in each task (see Figure 3):

- Binary (orange curve) peaks at 0.924 when k = 3 and stays above 0.90 through k ≈ 12, after which neighbour dilution sets in.
- Multiclass (blue curve) starts around 0.60 at k = 1, then slides toward 0.50 as larger neighbourhoods mix genres.

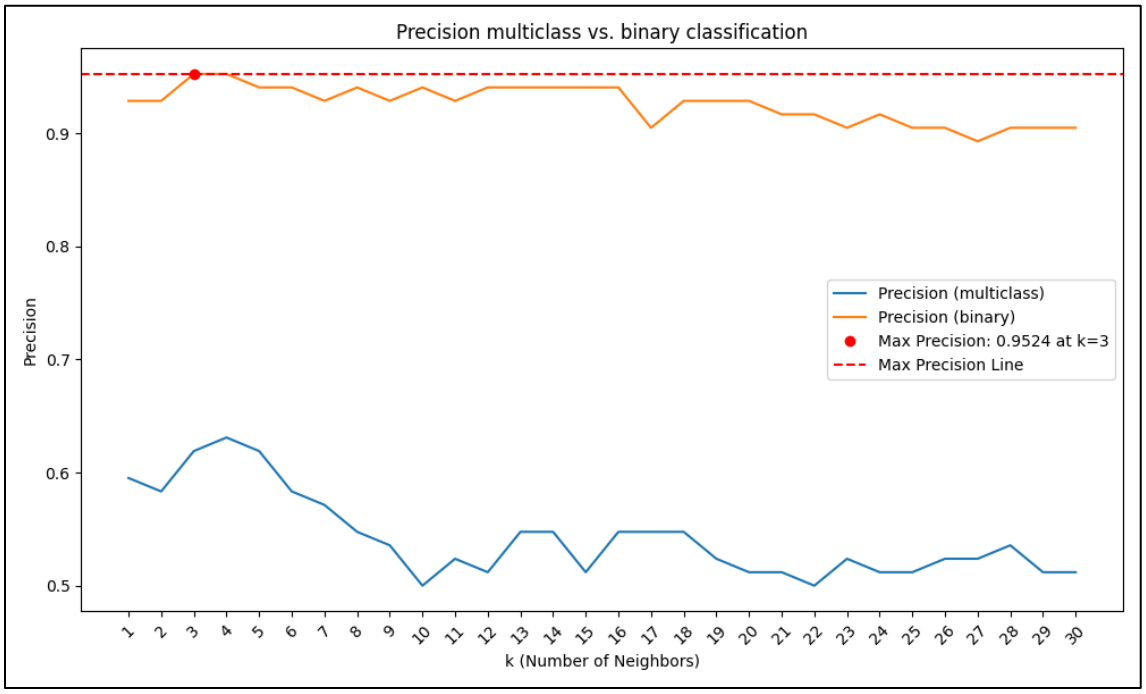

Figure 3. Precision as a function of k in multiclass and binary classification tasks.

Both versions of the final datasets were plotted using TSNE. See Figures 4, 5 and 6.

6.4 TSNE projections

Figure	Description
4	One 52-D vector per song
5	Three vectors per song
6	Binary view, three vectors

These figure-by-figure notes line up with the numbers in Section 5. Primus is unique enough to form its own cluster, but it still shares a little alt-metal and hard-rock DNA, so the model makes the odd mix-up.

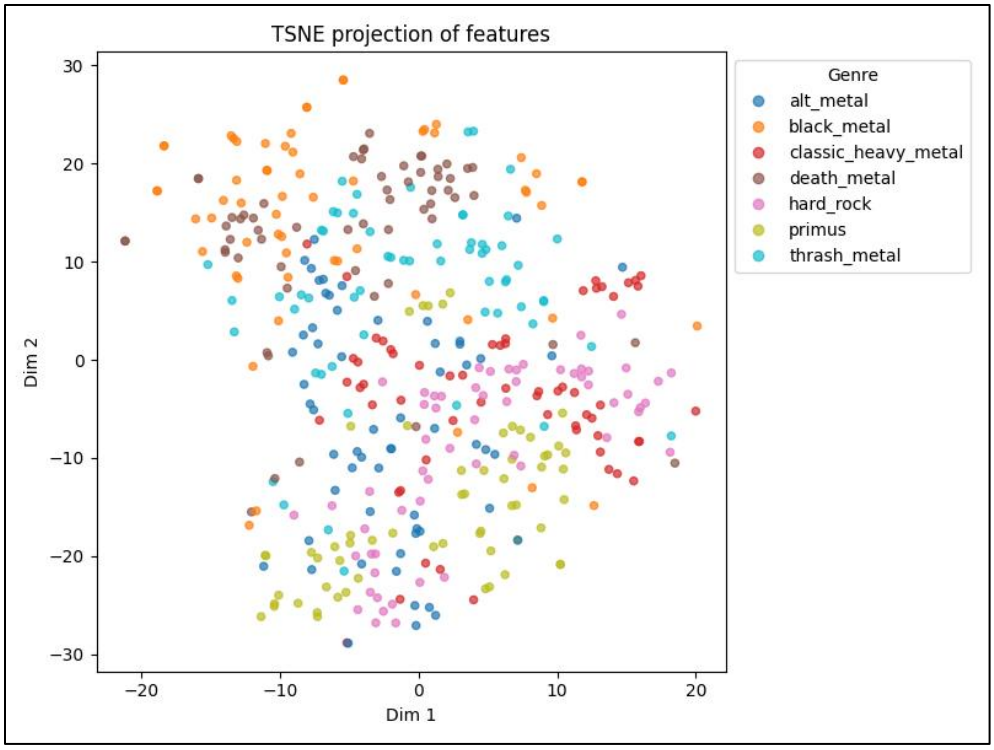

Figure 4. Primus forms a tight pink cluster just below the central axis; alt-metal (light blue) and hard-rock (brown) partially overlap on its flank.

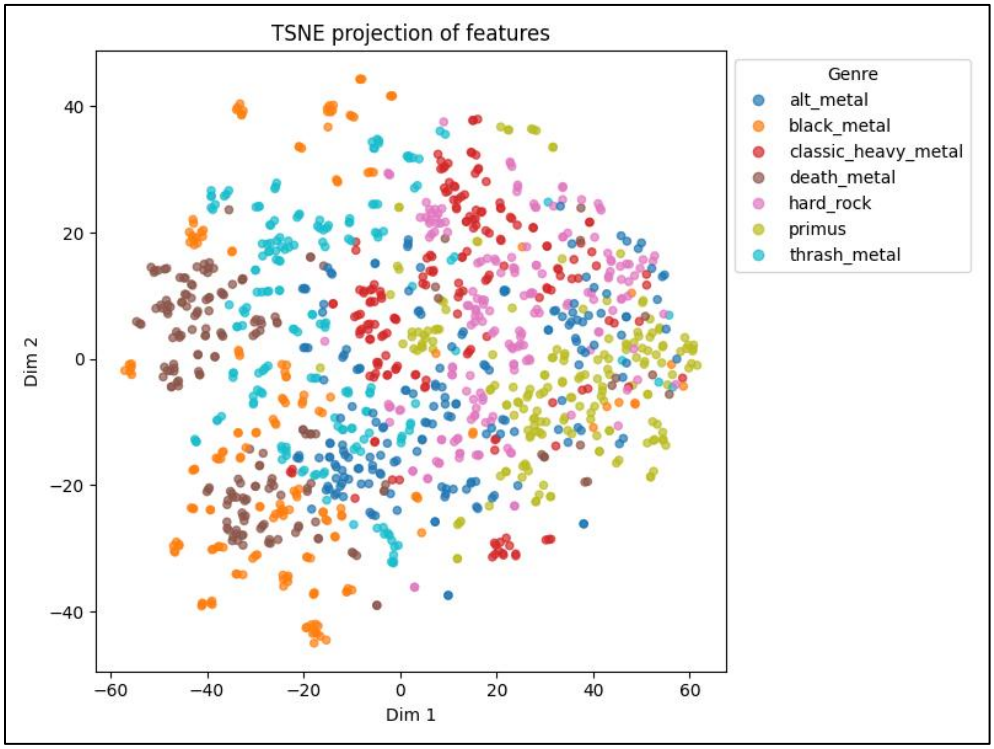

Figure 5. Splitting each track enhances separation: Primus drifts further from the hard-rock belt, showing that intra-song dynamics capture their quirkiness.

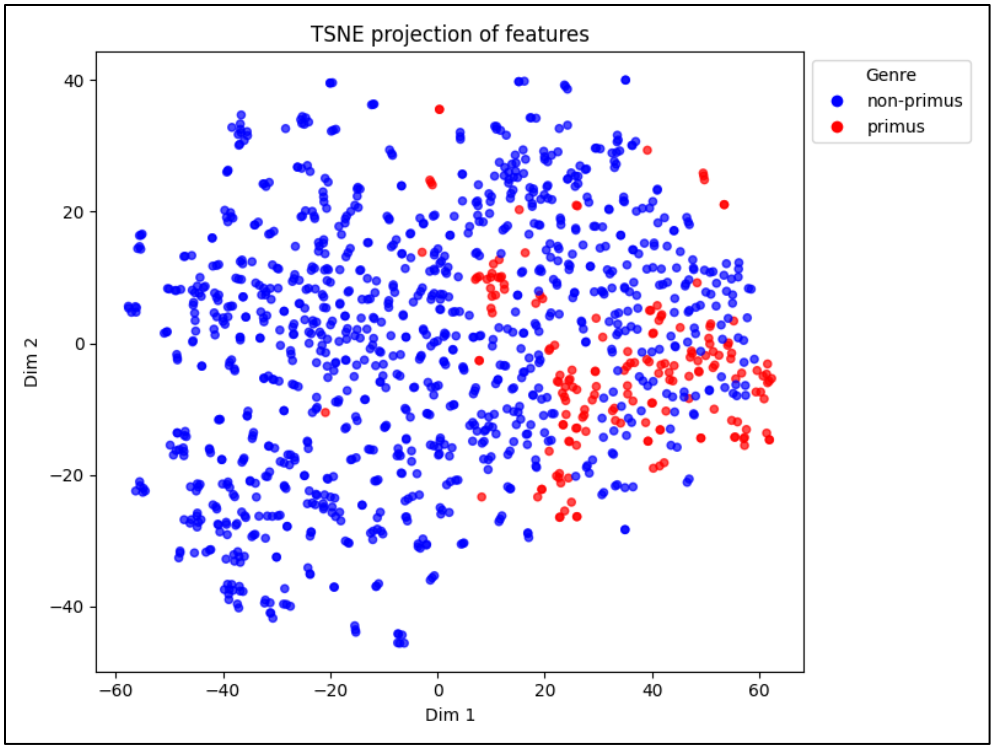

Figure 6. Primus points (red) are ring-fenced inside a broader blue cloud — visual proof of the 0.92 precision in Figure 2.

6.5 External validation on GTZAN

- **Figure 7** gives a 10 × 10 confusion matrix: classical, metal and blues top 90 % precision, while rock and country mingle (≤ 50 %).
- **Figure 8** echoes earlier trends — best macro-precision (0.675) appears at k = 2 and erodes steadily thereafter.
- **Figure 9** shows ten well-separated genre blobs under TSNE, confirming that the lower separation in our metal corpus is a property of the music, not the feature pipeline.

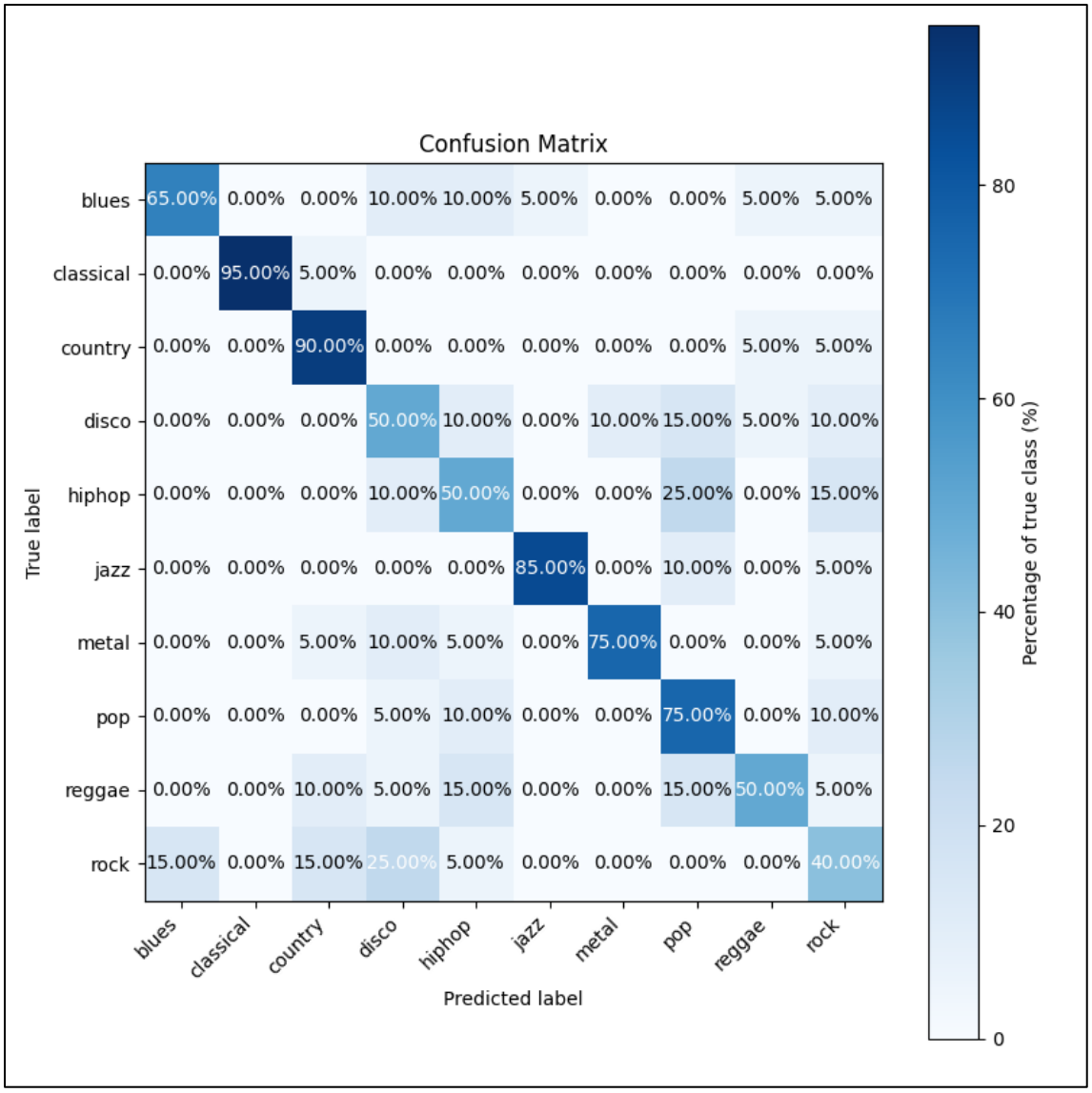

Figure 7. Confusion matrix of the external dataset

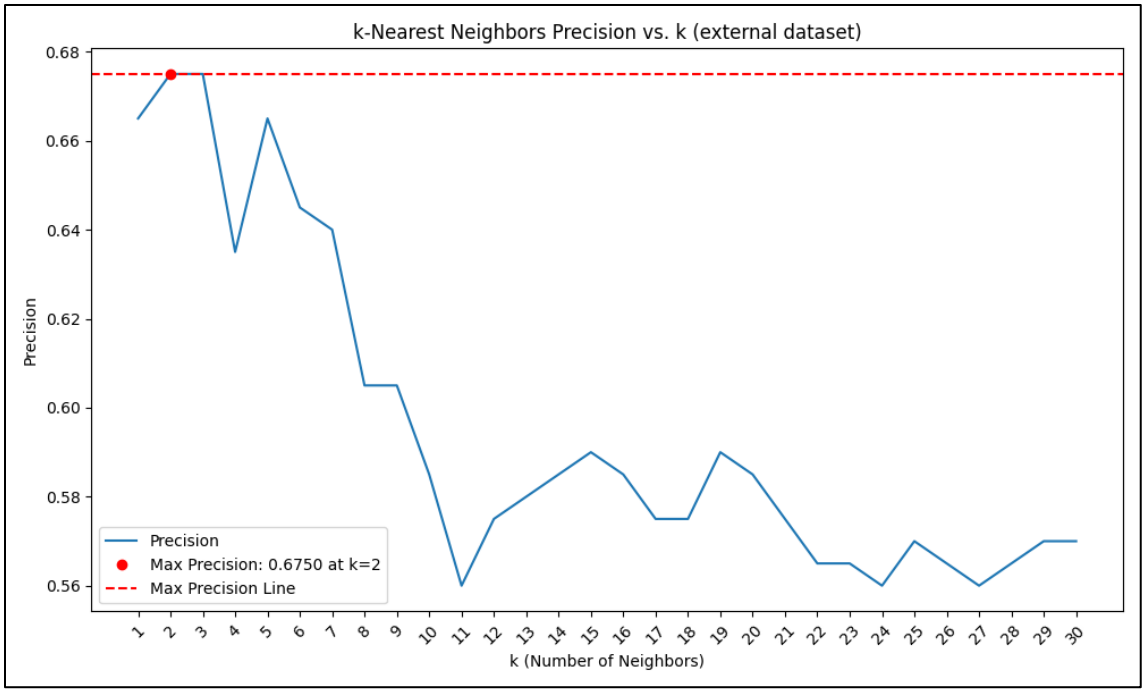

Figure 8. Precision as a function of k

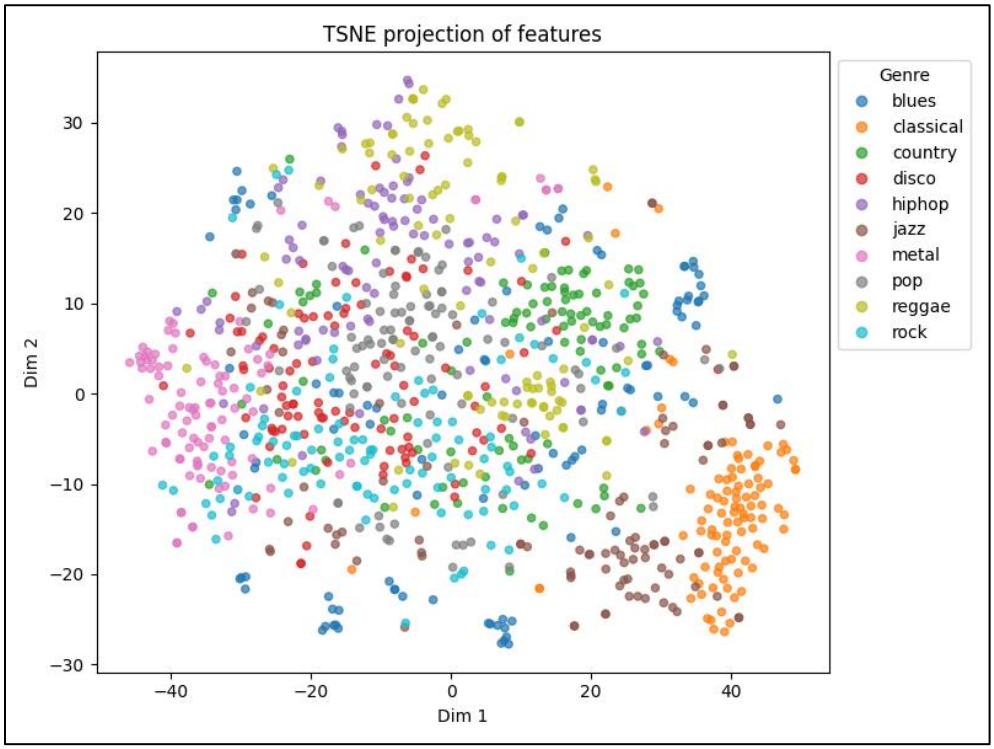

Figure 9. Projection of features using TSNE, external dataset

7 Discussion

7.1 What do the numbers tell us?

The best model scored 82 % in the six-way test and 92 % precision in the binary test, meaning Primus clearly has its own sound inside metal. A true-positive rate of 88.9 % for Primus — reached with only mean/variance MFCCs plus six standard MIR features — backs up the "genre-of-one" idea often heard from fans and journalists.

Key patterns in the confusion matrices:

- **Alt-metal & hard-rock mix-ups.** Most Primus mistakes are labelled as alt-metal or hard-rock. This fits comments that Les Claypool's slap bass feels like hard-rock, and many Primus riffs borrow from alt-metal rhythms.
- **Death- and black-metal stand apart.** Harsh vocals and fast tremolo guitars keep these styles far from Primus in feature space (no more than 5 % confusion). Extreme timbre still dominates MFCC-type features.
- **Segment-level boost.** When each track is cut into three equal parts, accuracy rises by about 6 percentage points. This means the model learns from the contrast between quirky sections and straight grooves — information that is lost when you average over the full song.

7.2 Musicological implications

Our numbers match the view of Primus as "carnivalesque metal." The band mixes odd time-signatures and sudden tempo jumps with unusual sounds like slap-bass and character-style vocals. Because Primus sits apart so clearly, our study supports the idea that genre lists should leave room for artist-level exceptions — single-entry tags — just as the ID3 standard already includes the special label PRIMUS.

7.3 Limitations

- **Dataset size & balance**. About 60 tracks per class is not large; a few wrong labels could move the scores by several points.
- **Manual genre tags**. Even experts disagree on sub-genre borders. A listener study could check whether the model's mistakes match human judgement.
- **Feature scope**. MFCCs capture timbre well but miss much of the rhythm and harmony. A CNN on mel-spectrograms or a joint timbre–tempo embedding might add useful detail.
- **Approximation error**. Annoy uses random projections, so it can return neighbors in the wrong order. Running an exact k-NN might change the thresholds a little.

8 Conclusion & Future Work

This project asked one clear question: Is Primus different enough to count as its own class inside metal?

Using a compact 52-dimensional feature vector — 20 MFCCs, six spectral-energy measures, and their means + variances over three song segments — and a fast k-nearest-neighbour model, we reached:

- 82 % overall accuracy in a seven-way sub-genre task, and
- 92 % precision for the "Primus" label in a binary Primus-versus-Rest set-up.

The confusion matrix shows Primus as a tight, machine-visible cluster. The few errors lean toward alt-metal and hard-rock, exactly the neighbors named by many music writers.

These findings back the idea of Primus as a "genre singleton" and prove that well-chosen summary features can reveal fine stylistic borders without heavy deep-learning models.

8.1 Future Work

- **Richer features.** Try mel-spectrogram CNNs or wav2vec-style embeddings to see if they push accuracy higher.
- **Perceptual validation.** Ask metal fans to label short clips and compare their errors with the model's.
- **Time-aware models.** Use sequence methods (e.g., GRU + attention) so the system can weigh different song sections.
- **Other outlier artists.** Apply the same pipeline to bands like Tool, Mr. Bungle, or Faith No More to test how general the method is.
- **Open resources.** Release code, feature tables, and split lists to make the study easy to reproduce and extend.

Acknowledgements

I am grateful to the open-source community — especially the teams behind the audio-analysis and machine-learning tools used here — for sharing their work so freely.

Warm thanks also go to my high-school friends, with whom I first explored the boundless landscape of music; their enthusiasm set the foundation for this project.

Finally, I acknowledge the musicians — Iron Maiden, Tool, Deftones, Primus, Swans, Radiohead, Red Hot Chili Peppers, Porcupine Tree, Opeth, Metallica, Pink Floyd and Korn — who continue to inspire me daily. Their creativity remains this study's deepest wellspring.

References

- Mulder, D.G.J. (2014). Automatic Classification of Heavy Metal Music. Bachelor's thesis in Mathematics and Computer Science.
- Rönnberg, T. (2020). *Automatic sub-genre classification of heavy-metal music.* (Master's thesis). Åbo Akademi University.
- Tsatsishvili, V. (2011). *Automatic subgenre classification of heavy metal music.* (Master's thesis). University Of Jyväskylä.
- Tzanetakis, G. (2002). *GTZAN genre collection [Data set].*
- Tzanetakis, G., & Cook, P. (2002). *Musical genre classification of audio signals.*