

### 1. Abstract

This project implements a music identification system that recognizes full songs from short audio segments (10 seconds) using Normalized Compression Distance (NCD) with multiple compression algorithms. The system extracts frequency representations from audio and compares them using different compressors (gzip, bzip2, lzma, zstd) to identify the most similar songs in a database. We evaluate the robustness of our approach by testing with various audio modifications including noise addition, pitch shifting, speed changes, reverb and equalization.

### 2. Introduction

Normalized Compression Distance (NCD) is a practical method to measure similarity between data objects using real-world compression algorithms. It approximates the theoretical Normalized Information Distance (NID), based on Kolmogorov complexity, the length of the shortest program that outputs a given string.

Since Kolmogorov complexity is non-computable, NCD is defined as:

$$\text{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Where:

C(x) and C(y) are the compressed sizes of x and y  
C(xy) is the compressed size of x and y concatenated

### 5. Future Work

#### Plagiarism Detection

Detect structural similarities to identify potential musical plagiarism

#### Copyright Monitoring on Streaming Platforms

Flag unauthorized or altered uploads, even with pitch or speed changes

#### Music Recommendation System

Recommend songs based on audio similarity for more perceptual results

### 3. Steps



Music files were converted into sequences of dominant frequencies over time



Short audio segments (~10s) were extracted to simulate queries



Some segments had noise added to test robustness

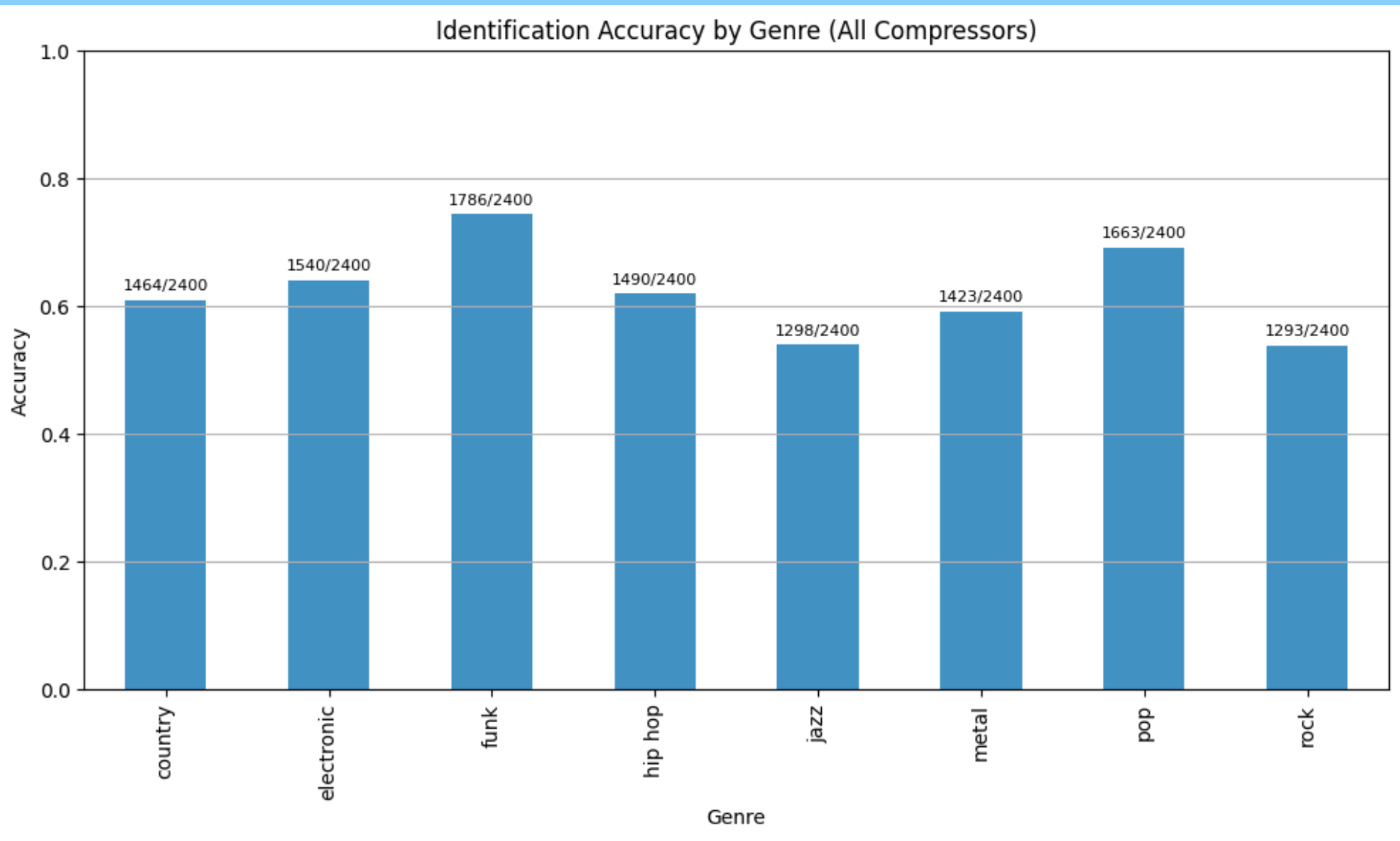
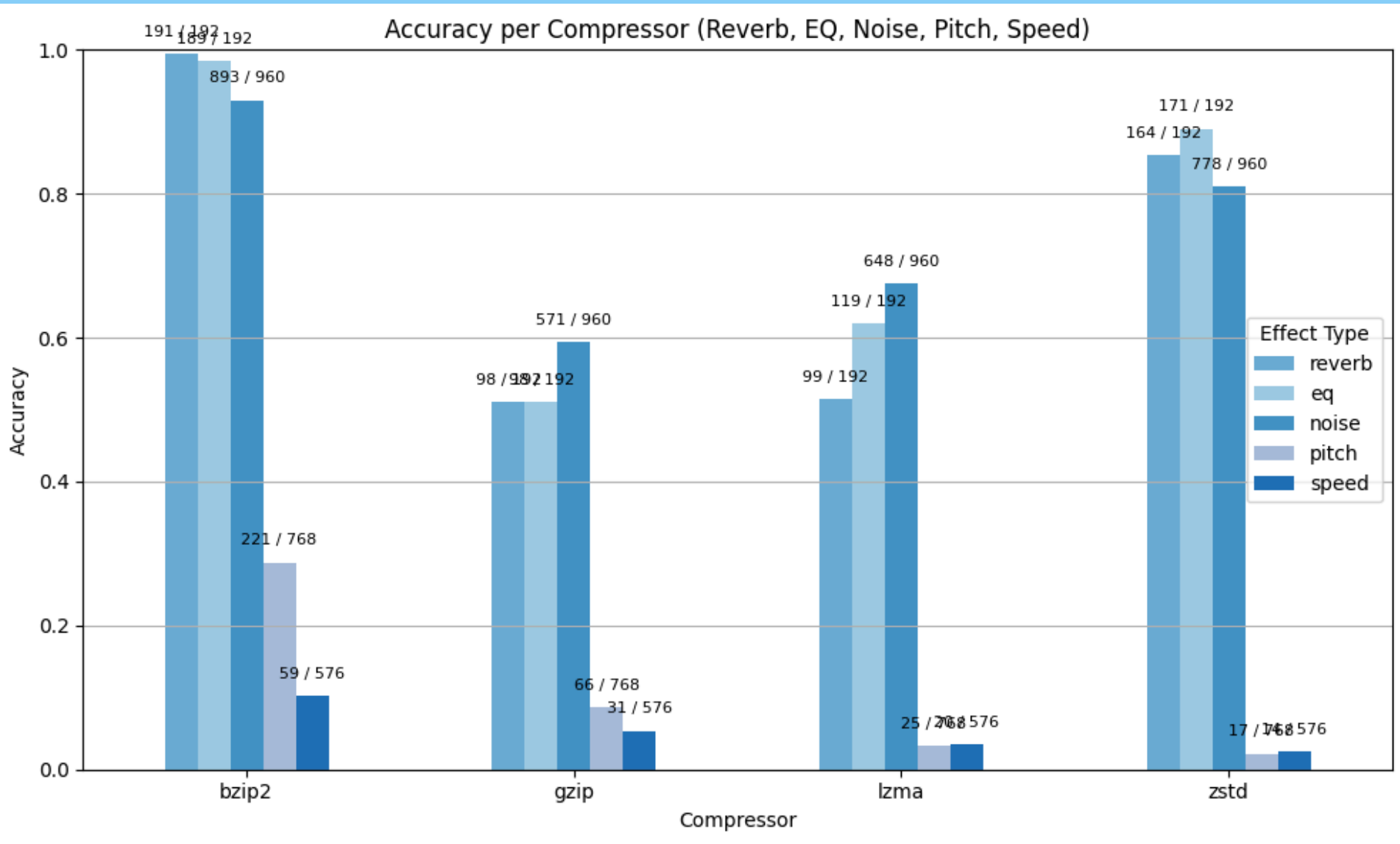
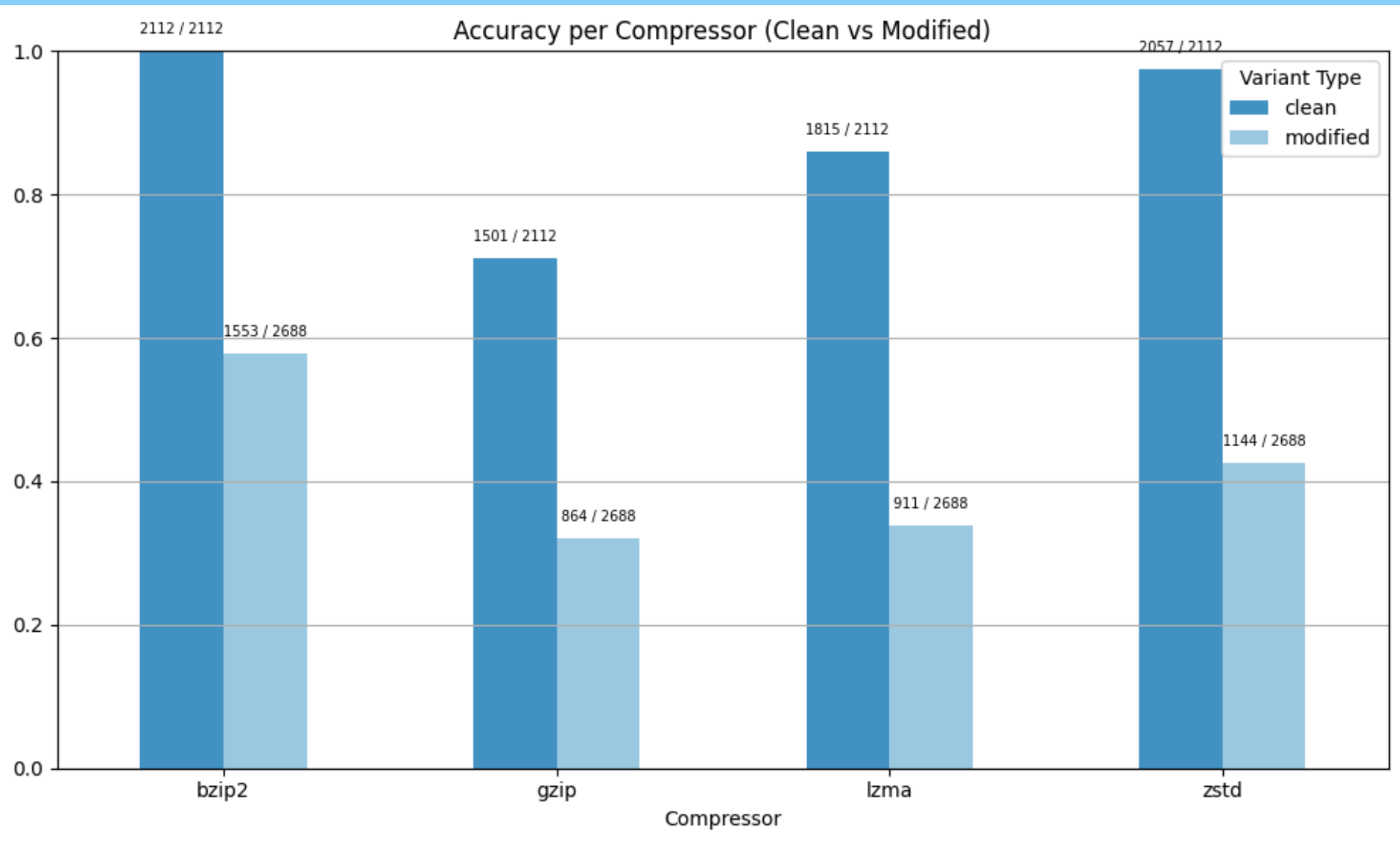


Segments were also converted into frequency-based representations



Segments were compared to full tracks using NCD. Compressors used: gzip, bzip2, lzma, zstd

### 4. Results



The best-performing compressor was bzip2, showing high accuracy in both clean and modified files. Around 20,000 different segment combinations were tested, covering various types of modifications, frequency counts, and window sizes. The analysis showed that speed and pitch changes had the greatest negative impact on detection accuracy. Interestingly, variations in the number of frequency bands and window size had little to no effect on overall performance. Among musical genres, funk was the easiest to detect when modified, while jazz was the most difficult.

### 6. Conclusion

This work explored the use of Normalized Compression Distance (NCD) with real-world compressors as a method for music identification using short audio segments. The approach is unsupervised, does not require manual feature extraction, and is inherently adaptable to various types of audio data.

The method demonstrated robustness to different kinds of modifications and offers a flexible, data-driven alternative to traditional audio analysis techniques. These characteristics make it a promising foundation for future developments in content-based music analysis and related tasks.