Datasheet Summary:

The combined dataset comprises approximately 1,700 musical pieces in .mp3 format sourced from NetEase music and 1000 audio tracks from the GTZAN dataset. The NetEase music dataset contains tracks with lengths ranging from 270 to 300 seconds and sampled at 22,050 Hz. These pieces are categorized into a total of 4 genres. The GTZAN dataset consists of 1000 audio tracks, each 30 seconds long, sampled at 22,050 Hz, covering four different music genres.

Data Fields:

- Classic/Opera
- Non_classic
  - Pop
  - R_and_b
  - Rock

Data Splits:

The combined dataset is commonly split into training and test sets, with 80% of the tracks used for training and 20% for testing. Care was taken to ensure an equal distribution of tracks from each genre in both the training and test sets.

Dataset Creation:

- **Curation Rationale:** Promoting the development of AI in the music industry.
- **Source Data:** Initial data collection and normalization by Zhaorui Liu and Monan Zhou.
- **Annotation Process:** Students collected about 1700 musical pieces (.mp3 format) with lengths of 270-300s divided into 4 genres.

Considerations for Using the Data:

- **Social Impact of Dataset:** Promoting the development of AI in the music industry.
- **Discussion of Biases:** Most are English songs.
- **Other Known Limitations:** Samples are not balanced enough.

Additional Information:

- **Dataset Curators:** Zijin Li

Data Usage (GTZAN):

- **Data Description:** The GTZAN dataset consists of 1000 audio tracks, each 30 seconds long, covering Blues, Classical, Pop, and Rock genres.
- **Data Collection Process:** Collected from various online sources and databases ensuring diversity.
- **Data Statistics:** Mean track duration: 30 seconds, Total tracks: 1000.
- **Data Cleaning and Preprocessing:** Minimal preprocessing including resampling and normalization.
- **Data Splits:** Commonly split into training and test sets with an equal distribution of tracks.
- **Metadata:** Each track accompanied by genre label and track ID, additional metadata not provided.

Citation

```
@dataset{zhaorui_liu_2021_5676893,
  author       = {Monan Zhou, Shenyang Xu, Zhaorui Liu, Zhaowen
Wang, Feng Yu, Wei Li and Baoqiang Han},
  title        = {CCMusic: an Open and Diverse Database for
Chinese and General Music Information Retrieval Research},
  month        = {mar},
  year         = {2024},
  publisher    = {HuggingFace},
  version      = {1.2},
  url          = {https://huggingface.co/ccmusic-database}
}

@misc{tzanetakis_essl_cook_2001,
author    = "Tzanetakis, George and Essl, Georg and Cook,
Perry",
title     = "Automatic Musical Genre Classification Of Audio
Signals",
url       = "http://ismir2001.ismir.net/pdf/tzanetakis.pdf",
```

publisher = "The International Society for Music Information Retrieval",
year   = "2001"
}

Datasheet Summary:

We created a dataset that comprised 16 musical pieces in .wav format soured form YouTube. The dataset contains tracks with lengths ranging from 250 to 300 seconds but cut into clips of 30 seconds long. The tracks are originally sampled at 22,050 Hz.

Data Fields:

- Experiment 1
    - -classic/opera
    - -pop
    - -r_n_b
    - -rock
- Experiment 2
    - -instrumental
        - -classic/opera
        - -pop
        - -r_n_b
        - -rock
    - -Acapella
        - -classic/opera
        - -pop
        - -r_n_b
        - -rock

Data Splits:

The combined dataset is not commonly split into training and test sets. 100% of the tracks were used for testing.

Dataset Creation:

- **Curation Rationale:** Promoting the testing of genre labels with mutated data

- **Source Data:** Initial data collection and normalization by Jatin Kulkarni, Anthony Ceja, Saiprathik Chalamkuri.
- **Annotation Process:** Students collected about 20 musical pieces (.wav format) with lengths of 270-300s divided into 4 genres before mutations.

Data Cleaning and Preprocessing: we performed speed-up and revered mutations to the 30-second clips we made for experiment 1. Then, for experiment 2, we found and split instrumental and acapella versions of our labeled data.

Citation

```
@datasheet{Cs342,
author    = "Jatin Kulkarni, Anthony Ceja, Saiprathik Chalamkuri",
title     = "Experiements dataset",
publisher = "CS342 Neural Nets",
year      = "2024"
}
```