Name: Frank (Runbao) Du
Faculty Supervisor: Prof. Blase Ur (Associate Professor, Dept. of Computer Science)

Area of Research: Characterizing the Artifacts of AI-generated Music versus Human-generated Music

Research Proposal

As AI's ability to perform tasks once thought to be uniquely human has rapidly expanded in recent years, AI generation has inevitably impacted the music industry, with AI music generation tools beginning to overwhelm the market. By 2028, generative AI music is projected to account for roughly 20% of music streaming revenues and 60% of music library revenues, with 24% of music creators' revenues at risk (CISAC 2024). As a music producer, I came across Suno during the process of making my own music, and was genuinely amazed by its ability to produce decent tracks. While exploring Suno, I was also introduced to the changes in the global music market prompted by AI-generated music, such as the destabilization of traditional creative workflows and the contested role of human artistry in an industry increasingly shaped by generative tools (Berger 2024).

These developments are exciting, but they also raise significant ethical questions about algorithmic monoculture and its implications for the music industry. If AI-generated music keeps becoming more and more prevalent, will we face a future where everything we listen to sounds similar? And when that happens, will we lose the humanness in human music? To even begin answering these questions, we first need to ask ourselves: what aspects of humanness are found in human music to begin with? What really composes human music, and what really composes music generated by AI? On a larger scale, can we use a quantitative measure to derive the humanness of human music? To address these questions, a deep analysis of the artifacts (unique traits of audio features) of AI-generated music becomes necessary. The goal of this research, beyond detecting artifacts and classifying audio files, is to examine the humanness within human music and to invite readers to critically reflect on what they listen to and what they appreciate. The methodology of this research consists of four parts, which will be explained in detail in the following sections.

Several past studies have tackled related questions. Afchar et al. (2024) use a controlled fake-generation setup: they autoencode real tracks from the FMA medium dataset using multiple decoder/codec pipelines so that semantic content remains constant and only generation artifacts vary. A small convolutional classifier trained on amplitude spectrograms reaches 99.8% accuracy in this setting. The work is foundational, but its limitations are significant. Detector accuracy collapses under common audio transformations such as pitch shift, EQ, and reverb; generalization across decoder families is weak; and the binary "real vs. fake" framing does not handle partially synthetic content well. The study also restricts itself to decoder/codec reconstruction artifacts, without engaging with higher-level semantic patterns such as genre conventions or compositional tendencies, or with the fully synthetic outputs of end-to-end platforms like Suno. Afchar et al. (2025) follow up on this line of work by asking what inherently makes synthetic audio detectable. They mathematically analyze the transposed convolution layers used in generative audio decoders and prove that these layers systematically induce periodization in the frequency spectrum, producing regular spectral peaks at frequencies determined by the stride configuration of the upsampling stack. A lightweight detector built on these "artifact fingerprints" achieves over 99% accuracy, rivaling much larger CNN and Transformer baselines. Its limitations mirror those of the

previous paper: the detector breaks under audio manipulations that alter frequency positions (resampling, pitch shift), and the analysis is confined to low-level signal artifacts produced by the decoder architecture, leaving open the question of whether higher-level musical features such as genre conventions, lyrical patterns, and instrumentation tendencies also carry detectable AI signatures.

My research is inspired by these papers in several ways, such as how they constructed their human-music datasets and the kinds of artifact fingerprints they identified in encoded audio. It also explores new territory by conducting research on fully synthetic output platforms like Suno, examining deeper semantic patterns such as lyrics and instrumentation, and aiming for a conclusion driven not by quantitative measures alone but by broader ethical and musical questions.

In Part 1 of the research, I will construct two datasets, one consisting of human-generated music and the other consisting of music generated by AI. The AI-generated music will be limited to output from Suno, as it is the leading AI music generation platform and transforms text prompts into complete songs with vocals, lyrics, and instrumental production (aicpb.com 2026). The human-generated music will be drawn from a combination of music archives such as FMA (Free Music Archive), MagnaTagATune, GTZAN, and MTG-Jamendo. These four archives are widely used as benchmark datasets in the music information retrieval community, which makes them a representative basis for analyzing human-generated music. FMA contains over 100,000 tracks across a wide range of genres, all released under Creative Commons licenses, and is commonly used in genre classification and audio analysis tasks. MagnaTagATune offers around 25,000 audio clips annotated with descriptive tags for mood, instrumentation, and style, providing metadata for cross-referencing musical features. GTZAN, although smaller with 1,000 tracks across 10 genres, remains one of the most established datasets for genre classification. MTG-Jamendo includes over 55,000 full tracks with detailed genre, instrument, and mood tags, offering strong coverage of contemporary independent music. Together, these archives provide both breadth and consistency, making them well-suited for constructing a representative human-generated dataset.

Both datasets will include audio files from eight different genres, taken from the Luminate 2025 End-of-Year Music Report as the most popular music genres of 2025. Classifying these songs by genre is crucial because tracks may exhibit different audio features across different genres. The goal is to maximize similarity within each of the human-generated and Suno-generated sets while minimizing similarity between them, so any other source of variation, such as that introduced by genre, must be controlled for. Due to differences in popularity across these eight genres, some have more available tracks than others. For now, I have decided to leave this imbalance as is rather than cutting down the number of tracks for popular genres in order to equalize counts, though this may change as the research progresses. The Suno-generated dataset will match the number of tracks in the human-generated dataset for each genre, mass-generated through API calls using only the genre name as the prompt. Both datasets will then be split into a training set and a testing set, with a ratio of roughly 7 to 3.

In **Part 2**, I will produce a set of musical data from the training set and conduct a detailed statistical analysis to identify the artifacts of Suno-generated music. I will first explore the possible measures, ranging from basic ones like BPM and key to more complex ones such as MFCC means and chroma entropy mean. During this process, I will begin by listing all candidate musical measures, then evaluate each in terms of its programmability and its relevance to this study. Relevance is determined through a small trial run on a batch of 200 Suno-generated songs alongside 200 human-generated songs: a measure

with high relevance should have a relatively consistent range within each of the two datasets but a different range across them, observable through simple statistics such as mean, median, standard deviation, and IQR. The selected measures will be attributed to one of the following groupings: dynamics, spectral, timbral, and rhythm. Each group will have its own Python pipeline that ingests music files one at a time and outputs a CSV containing all measures under that group across all input files. If a measure cannot be attributed to any group, it will have its own pipeline, such as the measure of lyrical structure, which does not fit cleanly under any of the four groups. After gathering data from the training set, a three-part statistical analysis will be conducted: first, summary statistics (mean, median, etc.) for each measure; second, two-sample tests (Welch's t-test or Mann–Whitney U depending on distribution shape, and chi-square for categorical measures such as key mode) to assess whether differences across the two datasets are statistically significant; and third, effect sizes (Cohen's d or rank-biserial correlation) to quantify the magnitude of each difference. Measures with both significant and substantively large effects will be flagged as candidate artifacts for the classifier constructed in the next part.

In **Part 3**, I will use the resulting catalog of artifacts to construct a classifier. The classifier should take audio and its corresponding genre as input, and output a numerical result indicating the degree of AI-generated content contained in the audio. The classifier will be intentionally simple and interpretable, built directly on top of the artifact catalog produced in Part 2. For each flagged measure, I will manually define a range of values associated with AI-generated traits, informed by the distributions and effect sizes computed in the previous step. Each measure will be assigned a weight reflecting its discriminative strength (higher weights for measures with larger effect sizes and tighter in-group consistency), and the classifier's output will be a weighted aggregate of how many of an input track's measures fall within their respective AI-trait ranges. This architecture sacrifices some raw performance compared to black-box models, but it preserves the ability to trace any classification decision back to specific artifacts. Optimization will proceed iteratively along several axes: tuning the per-measure weights and thresholds using cross-validation on the training set, conducting feature ablation studies to identify which measures contribute most and which may be redundant, and analyzing misclassified examples to reveal gaps in the artifact catalog that may warrant returning to Part 2.

Part 4 of the study explores Suno's prompting mechanism. The goal is to determine whether deliberate prompt engineering can successfully evade the classifier and degrade its accuracy. I will design several categories of prompt variations, each targeting a different aspect of the artifact catalog. The first category consists of artifact-aware prompts that explicitly attempt to suppress known AI traits, such as prompts requesting natural timing fluctuations, dynamic build-ups, or wider frequency content, in direct response to specific measures flagged in Part 2. The second category consists of genre-displacement prompts that push Suno outside of stereotypical genre conventions (e.g., requesting an acoustic folk arrangement under a hyperpop label). The third category consists of prompts that introduce irregular song forms, tempo changes, or unconventional instrumentation that human-generated music routinely exhibits. For each category, I will generate a fixed number of prompt variations, and multiple Suno generations for each variation. Each generated track will be passed through the classifier from Part 3, and its output score will be recorded. The quantitative analysis will compare these scores against a baseline distribution of scores from the testing set constructed in Part 1 to determine whether each prompt category produces a statistically meaningful drop in classifier accuracy. I will additionally examine which specific artifact measures shift most under each prompt category, revealing not only whether the classifier can be evaded

but also which artifacts are resistant to prompting and which are byproducts of typical prompting patterns and therefore evadable.

The study will conclude by reflecting on the broader ethical implications of AI-generated music and of building tools to detect it. While a classifier of this kind can support transparency and protect human artistry, it also carries risks of misuse, false positives, and an ongoing arms race with generation platforms, all of which warrant careful consideration alongside the technical findings.

Research Timeline
Spring 2026: Part 1
Summer 2026: Parts 2 and 3
Fall 2026: Part 4
Winter 2027: Manuscript preparation

References

Afchar, Darius, Gabriel Meseguer-Brocal, Kamil Akesbi, and Romain Hennequin. 2025. "A Fourier Explanation of AI-Music Artifacts." In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*. Daejeon, South Korea.

Afchar, Darius, Gabriel Meseguer-Brocal, and Romain Hennequin. 2025. "AI-Generated Music Detection and Its Challenges." In *ICASSP 2025: 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

AICPB. 2026. "AI Music Generator Rankings by Users — Feb 2026 Edition." AICPB, March 9, 2026. https://www.aicpb.com/ai-rankings/products/ai-music-generator-rankings.

Berger, Virginie. 2024. "AI's Impact on Music in 2025: Licensing, Creativity and Industry Survival." *Forbes*, December 30, 2024. https://www.forbes.com/sites/virginieberger/2024/12/30/ais-impact-on-music-in-2025-licensing-creativity-and-industry-survival/.

CISAC. 2024. "Global Economic Study Shows Human Creators' Future at Risk from Generative AI." International Confederation of Societies of Authors and Composers, December 4, 2024. https://www.cisac.org/Newsroom/news-releases/global-economic-study-shows-human-creators-future-risk-generative-ai.