Detecting AI-Generated Piano Music with a Spectrogram CNN: A Proof of Concept, and a Study in Shortcuts

Daniel Bordovský[1]

Abstract

We investigated whether a convolutional neural network can distinguish AI-generated piano music from human-performed piano recordings by treating the problem as image classification over mel-spectrograms. Using a deliberately narrow scope — solo and ambient piano only — we trained a ResNet-18 on 39 AI-generated tracks (Suno v5.5) and 42 human recordings (81 tracks, 3,914 three-second segments). The initial classifier reached near-perfect validation accuracy, however, subsequent analysis suggested that this performance may have been influenced by factors unrelated to musical content. In particular, we observed a systematic loudness difference between the two classes (mean RMS 0.127 vs. 0.098) and a hard high-frequency cutoff in the generated audio at ~18 kHz. After neutralising these confounds, imposing a common 16 kHz ceiling and matching loudness, in-distribution accuracy remained high (70.6%–99.2%, final 89.0%), and single-track inference labelled held-out tracks with very high confidence (≈90% for AI tracks, ≈95% for human tracks). One residual difference survived cleaning: a pronounced spectral-tilt (brightness) gap, with human recordings carrying markedly more energy across the ~1–9 kHz band. We retain rather than remove this difference, and argue it cannot be classified as confound or genuine artefact without an out-of-distribution test. The model failed to generalise: a track from a different generator (Udio) was misclassified, and two out-of-genre tracks (punk rock, one AI and one human) were both classified as "real" with full confidence. I additionally report qualitative observations from data collection — notably the generator's frequent non-compliance with prompts and a marked homogeneity of its piano timbre, and outline the data scale and the move beyond purely spectral features that a reliable detector would require. I conclude that a small, single-genre, single-generator detector is achievable and locally confident, but brittle and partially built on data-collection artefacts.

[1]Independent Researcher. Correspondence to: borrtex@borrtex.com

1. Introduction

The rapid improvement of generative music systems such as Suno and Udio has created a practical need to distinguish machine-generated audio from human recordings. A natural approach, borrowed from image forensics, is to render audio as a spectrogram and apply a standard image classifier, on the hypothesis that generative models leave characteristic artefacts — in harmonic structure, transients, or noise floor — that survive as visual patterns in the time–frequency representation.

This article documents a proof-of-concept built around that hypothesis, undertaken as an exploratory single-operator experiment rather than a formal study. We restricted the problem to a single instrument and texture: solo and ambient piano, on the reasoning that a narrow domain makes any genuine generative artefact easier to isolate and makes confounds easier to detect. The experiment succeeded narrowly and failed broadly, and both outcomes are informative.

2. Methods

2.1 Data

AI class. 39 instrumental piano tracks were generated with Suno v5.5 (Pro tier). To sample the generator's output distribution rather than a single prompt's idiosyncrasies, we used a rotating set of prompts varying in sub-style, tempo, and recording character. Short style prompts included, for example:

- solo classical piano, slow, expressive rubato, intimate, close-mic
- ambient piano, sparse notes, long sustain, soft drone underneath, reverb
- cinematic solo piano, building intensity, emotional

We also tried longer, highly specified prompts intended to elicit a more authentic, less produced tone, for example: *"intimate contemporary piano piece, minimalist composition built from repeating motifs and gentle evolving patterns, spacious emotional atmosphere, delicate touch, subtle imperfections, expressive dynamics, natural pedal resonance, soft hammer noise, warm felted tone."* Style exclusions (vocals, strings, drums, guitar, synth leads) were applied to enforce a piano-only constraint.

Human class. 42 solo piano recordings assembled from real acoustic sources, including original works by the author (a composer) and by collaborators, who are not individually identified here, together with a broader selection of piano tracks. The material spanned ambient grand-piano pieces and intimate felted, close-miked piano. All of it originated from real microphone-captured piano sound, whether recorded as full live performances or produced with sampled virtual instruments (sample libraries built from microphone recordings of real pianos). None of it was AI-generated, and none used purely synthetic or generic tone generators.

This implies a deliberate definitional choice that we revisit in Section 7: the contrast studied is generative-AI-synthesised audio versus real-acoustic-sourced audio (which includes sample-library productions), rather than "live human performance" versus "everything else." In total the corpus comprised 81 unique tracks.

2.2 Qualitative observations during generation

Two observations from the generation process are worth recording, as they bear on both the difficulty of building the dataset and the interpretation of the results.

First, the generator frequently did not respect the prompt. Requests for solo piano jazz, for instance, almost always returned a track with added drums, despite explicit instrumentation constraints; the longer, more detailed prompts were not noticeably better respected than the short ones.

Second, and more interesting, the generated piano was timbrally homogeneous. Across prompts, the tone and character of the instrument were very similar — consistently close to a bright, well-produced grand piano — and it proved nearly impossible to push the output toward a more idiosyncratic or authentic timbre (e.g. an upright or felted piano). A plausible explanation is the composition of the generator's training data: piano recordings online are dominated by grand pianos, so the model appears to collapse toward that modal timbre regardless of the prompt. This homogeneity is notable because it parallels the low loudness variance we later measured in the AI class (Section 3) and may contribute to the residual brightness signature (Section 4): a generator that always produces the same well-produced grand-piano tone yields a class that is unusually consistent, and consistency is itself a thing a classifier can learn.

2.3 Preprocessing and segmentation

Each track was decoded to mono and divided into non-overlapping 3-second segments. Each segment was converted to a mel-spectrogram (128 mel bands, maximum frequency 16 kHz),

expressed in decibels with per-segment normalisation to the segment maximum, and rendered as a fixed-size image. This yielded 3,914 spectrogram images.

The train/validation split was performed at the track level, not the segment level: all segments from a given track were assigned wholly to either the training or validation set (64 tracks / 3,184 images for training; 17 tracks / 730 images for validation). This prevents the most common form of leakage in audio classification, in which segments of the same recording appear on both sides of the split and the model is rewarded for recognising the *song* rather than the *class*.

2.4 Model and training

We fine-tuned an ImageNet-pretrained ResNet-18, replacing the final fully connected layer with a two-class output. Inputs were resized to 224×224 and normalised with ImageNet statistics. Training used the Adam optimiser (learning rate 1×10^{-3}), cross-entropy loss, and ran for 10 epochs on a GPU.

2.5 Evaluation

We report per-segment validation accuracy during training, and for single-track inference we aggregate per-segment softmax probabilities across all 3-second windows of a track into a single track-level probability.

3. Initial Results and the Confound Problem

The first trained model reached validation accuracy in the mid-to-high 90s, occasionally touching 100%. For a 3-second clip of solo piano, this is implausibly good, and was treated as a warning sign rather than a success.

Having ruled out track-level leakage by construction (Section 2.3), we examined whether the two classes differed in some global property unrelated to musical content. They did, in two ways (Figure 1).

(1) Loudness. The generated tracks were systematically louder than the human recordings (mean RMS 0.127 vs 0.098) and, tellingly, far more uniform in loudness (standard deviation 0.027 vs 0.039). This is consistent with a generator that applies consistent automatic mastering to every export, whereas human recordings from many artists vary in their mastering.

(2) Bandwidth. The mean power spectrum revealed a hard, brick-wall low-pass in the generated audio at approximately 18 kHz, above which energy collapsed by tens of decibels, while the human recordings retained energy up to roughly 20–22 kHz. Quantitatively, the 99.5%-energy spectral roll-off differed by 4.77 kHz (AI 5.66 kHz vs human 10.43 kHz). Because both classes were stored as high-bitrate MP3, this cutoff is a property of the generated signal, not of the file container. A classifier can separate the two classes almost perfectly by asking a single question "is there energy above 18 kHz?" without learning anything about pianos.

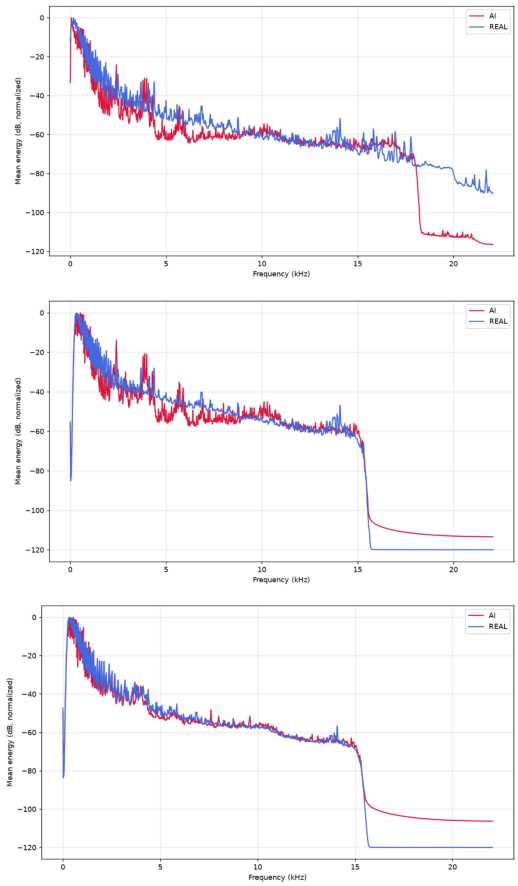


Figure 1.

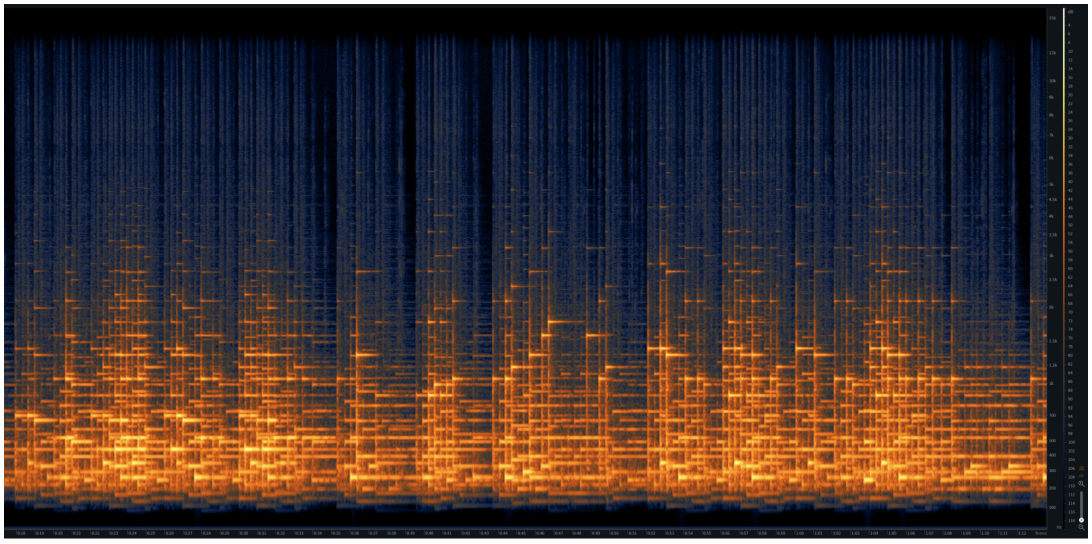
Top: Mean spectrum before any processing. The AI tracks (red) fall off sharply at ~18 kHz, while the human tracks (blue) carry energy to ~20–22 kHz, and the two differ in spectral tilt across the band.

Middle: Mean spectrum after cleaning. Both classes now fall off together at the shared 16 kHz ceiling, and loudness is matched. The blatant cutoff and loudness shortcuts are gone. (Energy above the cutoff is numerical/resampling residue below the 16-bit floor and outside the model's input range.)

Bottom: Mean spectrum after additionally whitening the spectral envelope, so each class's average spectrum is forced to match. The red and blue curves now almost coincide across the whole band. This is achievable, but normalisation can always be pushed until the two classes are indistinguishable, at which point the detector is left with nothing to learn. We stopped at loudness and the shared 16 kHz ceiling and kept the residual spectral-tilt difference, rather than over-polishing it away as shown here.

These findings frame the central question of the project, which recurs throughout: *is a given class difference a confound or a genuine artefact?* A confound is an accident of how the data were collected (here, the generator's mastering and codec history) that will not generalise. A genuine artefact is an intrinsic property of the generative process that a detector should legitimately exploit. The two are not distinguishable by inspection; they must be separated empirically.

To complement the *averaged* spectra, it is worth looking at a single matched pair of tracks at full resolution, one AI (Suno v5.5) and one human recording of comparable piano material (Figure 2). What appears in the mean spectrum as a statistical tendency is visible here in a single track: the AI example is relatively *scooped* in the ~5–8 kHz mid-band and *lifted* in the ~8–12 kHz region, while the human recording tapers more smoothly and continuously toward the high end. This is the same signature the mean-spectrum difference (Figure 1) reported — human material carrying more energy through the mids, it is a plausible cue the CNN latched onto, since it is consistent and easy to read off a spectrogram.



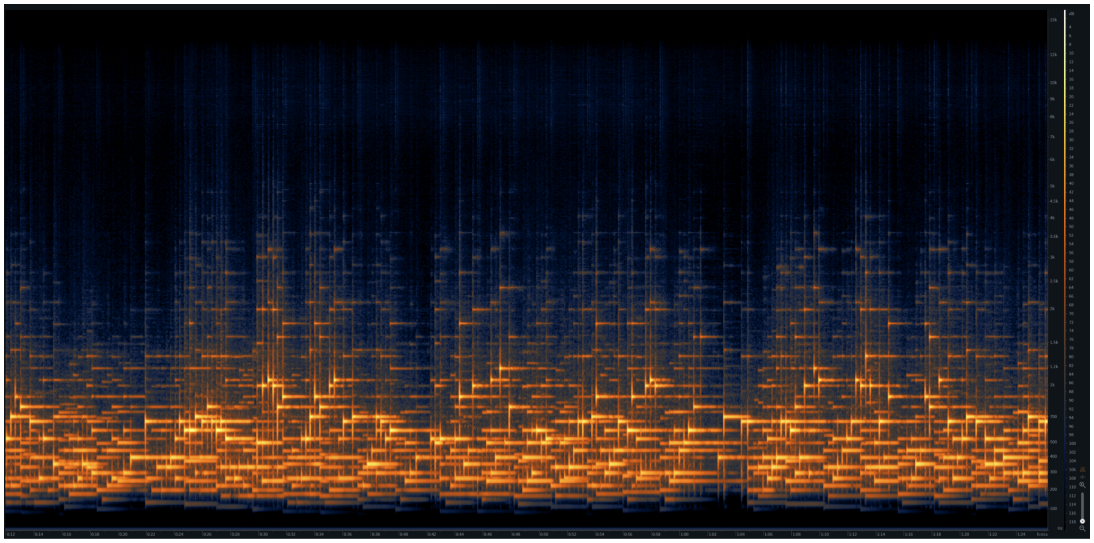


Figure 2. **Top:** Full-resolution spectrogram of a comparable human piano recording. Energy tapers smoothly and continuously toward the high frequencies, without the mid-band scoop and upper-band lift seen in the AI track. **Bottom:** Full-resolution spectrogram of an AI track (Suno v5.5), log-frequency axis. The mid-band around ~5–8 kHz is relatively suppressed while the ~8–12 kHz region is comparatively lifted, an EQ-like signature that matches the averaged measurement of Figure 1.

Two caveats apply, both familiar from §4. First, this is a single pair, illustrative rather than statistical. Second, the cause is ambiguous: an EQ-like mid-scoop-and-upper-lift could be a deliberate, consistent mastering choice in the generator's output as readily as a synthesis artefact, the intrinsic-versus-incidental question again. What makes it a *usable* cue is precisely its consistency across both the averaged statistic and the individual example; what it is not, on this evidence alone, is proof of a generalisable "AI" signature.

4. Confound Mitigation

We processed both classes through a single, identical pipeline so that neither could carry a pipeline-specific signature:

1. **Common frequency ceiling.** All audio was resampled to 32 kHz, imposing a shared 16 kHz Nyquist limit. This removes the generator's 18 kHz cutoff as a discriminative feature and forces the model into the 0–16 kHz band both classes share.

2. **Loudness normalisation.** All audio was normalised to a common loudness target, eliminating both the level difference and its suspicious uniformity.

3. **Low-frequency control.** A 250 Hz high-pass was additionally applied to both classes; retraining with it left accuracy essentially unchanged, indicating the low band was not a primary basis for the model's decisions.

After cleaning, the two classes were matched on the cues we had identified: the loudness difference fell to 0.0004 (from 0.0289), and the spectra overlapped above the now-shared 16 kHz ceiling.

One difference survived, however. The 99.5%-energy roll-off still differed by 2.10 kHz (AI 8.31 kHz vs human 10.41 kHz): a difference of *spectral tilt*, or brightness, rather than a hard cutoff. A difference curve between the smoothed class spectra makes its magnitude and location explicit (Figure 3): the human recordings carry substantially more energy than the AI tracks across the entire ~1–9 kHz band, by up to ~13 dB, with the AI tracks only marginally brighter in narrow patches around 9–11 kHz and 14–15 kHz.

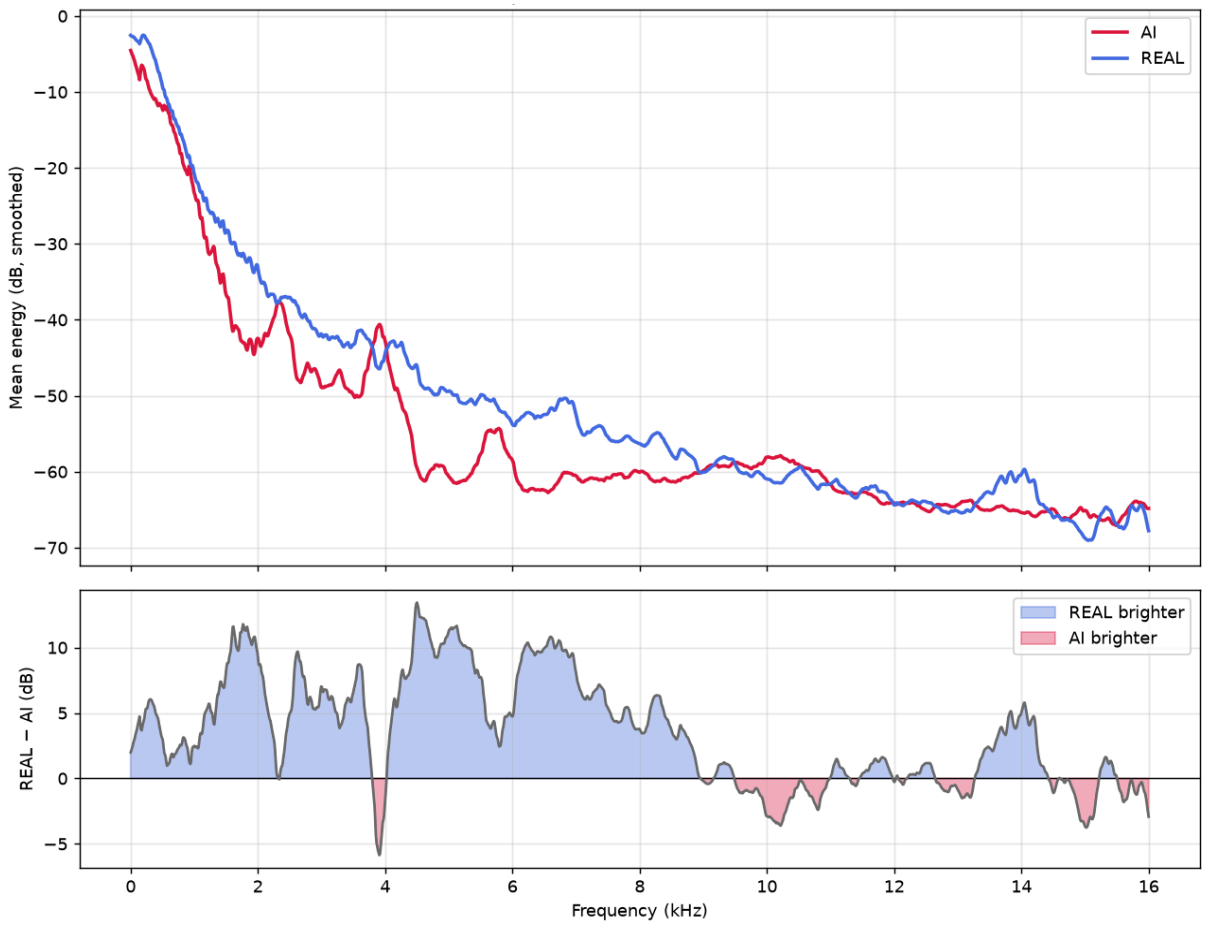


Figure 3. Smoothed mean spectra (top) and the human-minus-AI difference (bottom) after cleaning. Blue regions mark where human recordings carry more energy; the human class is markedly brighter across ~1–9 kHz. This residual brightness gap is the most likely basis for the model's post-cleaning decisions.

We deliberately did not remove this brightness difference. Unlike loudness and the codec-level cutoff, a difference in mid-to-high-frequency content is plausibly *intrinsic*: a generative model that fails to reproduce the bright hammer-attack transients and upper partials of a real piano, or one that, as observed in Section 2.2, collapses to a single homogeneous timbre, would produce exactly this kind of signature, and detecting it would be legitimate. But it is equally consistent with the human recordings simply being mastered brighter than Suno's output. The roll-off statistic and the difference curve are identical in both cases: they show the *size* of the gap, not its *cause*. This illustrates a general principle that bounded the cleaning effort:

two sources are never identical, so a spectrum will always differ at sufficient magnification; one cannot normalise away every difference, because in the limit of identical distributions the classes become undetectable by construction.

Whether this particular difference is signal or confound cannot be decided from the spectrum; it requires an out-of-distribution test (Section 6).

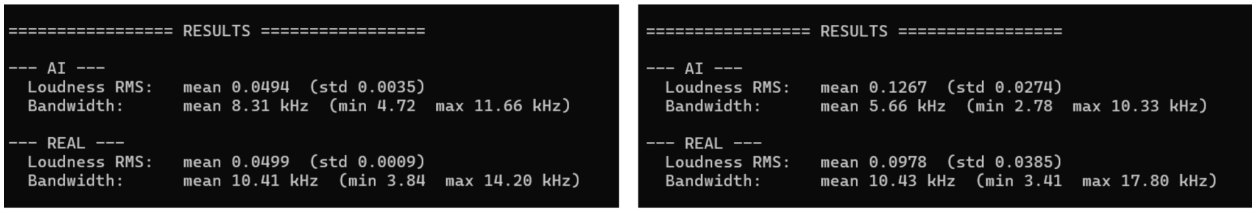


Figure 4. **Left:** Before cleaning. The AI class is louder (RMS 0.127 vs 0.098) and more uniform in loudness, and rolls off much lower (bandwidth 5.66 vs 10.43 kHz) — the measured form of the confounds seen in the spectrum. **Right:** After cleaning. Loudness is now matched (0.049 vs 0.050) and the bandwidth gap has fallen from 4.77 to 2.10 kHz. Only the residual brightness difference remains.

After all cleaning, validation accuracy remained high but *unstable*, ranging from 70.6% to 99.2% across the ten epochs (final epoch 89.0%; training accuracy ~99.8%). This instability is itself informative: a trivial shortcut produces stable, near-perfect accuracy, whereas the observed variance, on a validation set of only 17 tracks is consistent with a model relying on a feature present in most but not all tracks. It also indicates that single-number accuracy on a validation set this small should be interpreted cautiously.

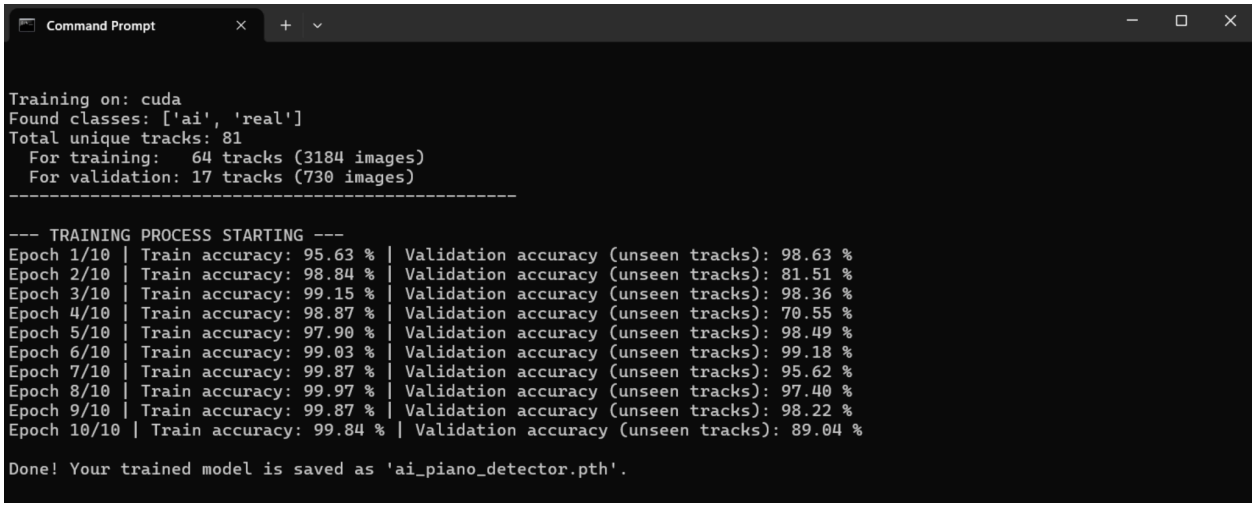


Figure 5. Training the detector after cleaning (81 tracks, track-level split). Validation accuracy is high but unstable across epochs (70.6%–99.2%, final 89.0%).

5. Behaviour at Inference, and Generalisation Failures

We built a track-level inference tool applying the identical preprocessing chain and aggregating per-window probabilities.

In-distribution performance was very good. A held-out Suno v5.5 track was classified as AI with ≈91.0% confidence; a held-out human recording was classified as real with ≈97.3% confidence.

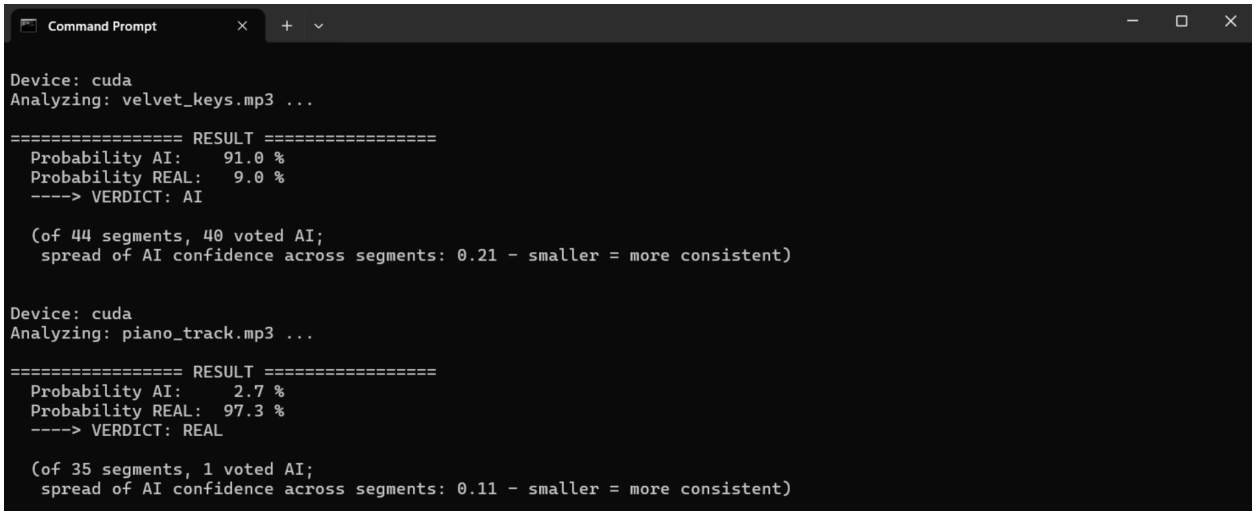


Figure 6. Single-track inference and how the verdict is formed. Each track is sliced into 3-second segments, every segment is scored independently by the model, and the per-segment probabilities are averaged into one track-level verdict; the number of segments voting "AI" and the spread of their confidence show how consistent that decision is. Here an AI track (velvet_keys) is called AI at 91%, 40 of its 44 segments agree, while a human recording (piano_track) is called real at 97.3%, with just 1 of 35 segments dissenting. The low spreads (0.21 and 0.11) indicate the verdict is stable across the whole track rather than driven by a few windows.

The model's limitations appeared as soon as it was taken outside its training conditions.

Unseen generator (Udio). A track from Udio, a generator entirely absent from the training data, was handled poorly. This result suggests that the detector may have learned characteristics specific to the generator represented in the training set rather than features that generalize across AI-generated music more broadly.

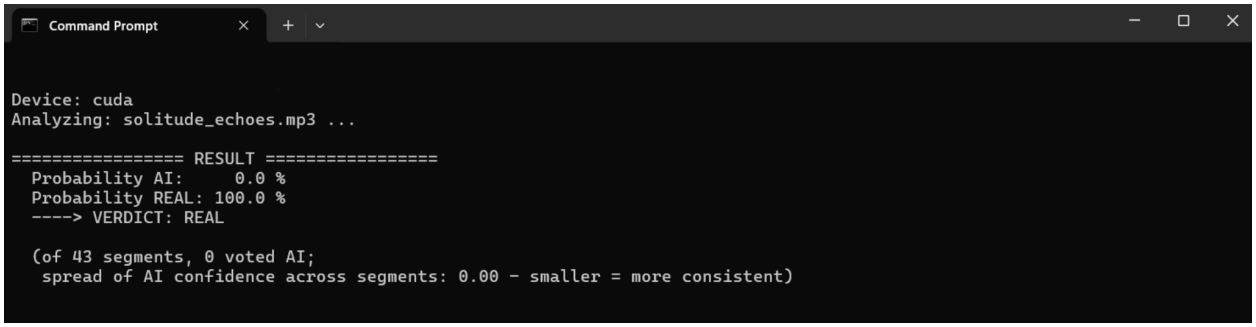


Figure 7. Classifier output for the Udio generated track, illustrating a case where the model failed to recognize audio from an unseen generator.

Out-of-distribution genre (punk rock). Presented with two punk rock tracks, one AI-generated and one human, the model classified *both* as "real" with 100% confidence. The interpretation is instructive, the model never learned the spectral texture of dense, high-energy guitar music, so neither track resembled the AI-piano patterns it had internalised, and it defaulted to "real." This is a textbook out-of-distribution failure, and a dangerous one for deployment, because the failure mode of an unfamiliar input is a confident *false negative*.

Spectral density and energy. These failures also point to a content-dependent difficulty. Low-energy, sparse material such as solo piano leaves much of the time–frequency plane empty, exposing subtle generative artefacts. High-energy, spectrally dense genres (techno, metal) fill the

spectrum and would likely mask the same nuances, making detection harder precisely in the genres where generated music is increasingly common.

6. Discussion

The experiment supports a narrow positive claim and a broad negative one.

The narrow positive claim is that a modest dataset (~40 tracks per class) suffices to train a classifier that distinguishes one generator's output from human recordings, *within a single genre*, with very high in-distribution confidence. To determine whether a track was generated by Suno or is a genuine human piano recording, the proof of concept works.

The broad negative claim is that this confidence does not constitute a general AI-music detector, for several compounding reasons:

- **Generator specificity.** The model learned features of one generator (Suno v5.5) and did not transfer to another (Udio).
- **Genre specificity.** The model learned features of one genre and collapsed to a default class outside it (punk rock).
- **Fragility to generator change.** The most reliable features identified, mastering and high-frequency behaviour, are exactly the properties a vendor can change at will.
- **Confound entanglement.** Some discriminative signal originated in data-collection artefacts (loudness, codec bandwidth, bitrate) rather than the generative process. We removed the most blatant of these, still the residual brightness gap remains unclassified.

The single experiment that would resolve the residual brightness gap, and the project's central ambiguity, is a generalisation test: evaluating the trained model on human piano from a wholly independent source (a different recording and mastering lineage it cannot have memorised). If accuracy holds there, the brightness it reads is intrinsic to AI-versus-human and is a legitimate artefact; if it collapses, the gap was the production style of this particular human corpus.

6.1 Toward a reliable detector

Scaling this proof of concept into something dependable would require, at minimum, far more data, on the order of a thousand tracks per class rather than forty, and almost certainly a *partitioned* training strategy: a separate model (or a conditioning signal) per genre, and a separate model per generator, since each genre has its own spectral texture and each generator its own fingerprint. Three problems follow directly from this.

(1) The human side is the hard side to collect. Assembling thousands of genuinely human recordings, across genres, with the legal rights and consent to train on them, is a substantial undertaking, arguably harder than collecting AI examples, which can be generated on demand. A detector is only as good as the breadth and legitimacy of its "real" corpus.

(2) The target moves. Even a detector trained on every current genre and every current generator could be undermined by a single model update or new version. A vendor changing its mastering, its frequency handling, or its underlying architecture can silently shift the very features the detector relies on, without changing the music's perceptual "AI-ness."

(3) Spectral analysis has a ceiling. Today's generators leave low-level spectral fingerprints, band-limiting, mastering uniformity, timbral homogeneity, that a spectrogram CNN can exploit. But as systems improve, they will increasingly reproduce the surface characteristics of real recordings, including microphone noise, room tone, and natural high-frequency content, eroding those fingerprints. At that point a purely spectral approach is insufficient. The natural response is a model that reasons about higher-level musical structure rather than surface texture alone: tempo and micro-timing, the logic and evolution of chord progressions, how individual instruments develop and interact, and the cohesion and balance of a full arrangement over time. These are properties where current generators are weaker and which are harder to fake convincingly than a noise floor.

This points to a difference between audio and image forensics that is worth stating carefully. Image-generation detection does not, in practice, recover the literal "layers" a generator stacked; it exploits low-level statistical fingerprints left by the synthesis process — upsampling and convolution artefacts, periodic frequency-domain traces, diffusion or GAN residuals. The spectral artefacts exploited in this study are the direct audio analogue of those fingerprints, so the two modalities are more similar than they first appear. The divergence is in where each must go *next*: as fingerprints vanish, music offers an unusually rich higher-level structure — temporal, harmonic, and multi-instrument — for a detector to interrogate, a dimension with no clean equivalent in a single still image. A future audio detector will likely have to listen the way a musician listens, not simply look at a spectrogram the way an image classifier looks at pixels.

---

7. Limitations

This was an exploratory single-operator experiment and should be read as such. The dataset is small (81 tracks) and covers one genre, one primary generator, and one generator version. The human and AI corpora were not matched on recording provenance, so the residual brightness difference may reflect production style rather than synthesis. The human class, by design, includes also sample-library (VST) productions of real microphone-recorded pianos; the study therefore distinguishes generative-AI synthesis from real-acoustic sources, not live performance from programmed performance. The validation set (17 tracks) is small enough that per-epoch accuracy is noisy (Section 4), and training is not fully deterministic, so reported figures correspond to a representative run rather than a fixed seed. We did not carry out the systematic external test against an independent human-piano corpus that would be required to estimate a true false-positive rate and to resolve the brightness ambiguity. The generalisation observations (Udio, punk rock) are single examples, illustrative rather than statistical.

8. Conclusion

A spectrogram-plus-CNN pipeline can learn to identify AI-generated piano from a small dataset and report high confidence on in-distribution tracks. That success is real but local: it is specific to one generator and one genre, partially dependent on data-collection artefacts, and fragile to both generator updates and out-of-distribution inputs, where it fails by confidently defaulting to "real." Building a detector trustworthy in the wild is a substantially larger undertaking — broad in genre, deep in data, comprehensive across generators, robust to encoding, and ultimately reaching beyond spectral texture into musical structure. The lasting value of this proof of concept lies less in the working classifier than in the demonstration of how a model can be right for the wrong reasons, and how to find out.

Notes. Implementation used librosa for audio analysis and spectrogram generation, PyTorch/torchvision for the ResNet-18 model, and scipy for filtering. Generated audio was produced with Suno v5.5 (Pro); the cross-generator test used a single Udio track. All figures and accuracy values come from a single small-scale run and are reported as approximate.