 ISSN (Online): 3140-5312

Distinguishing Professional and Amateur Piano Performances using 2D Audio Features and Convolutional Neural Networks

Authors and affiliations:

Changheun Oh*, Department of Software and Communication Engineering, Hongik University, Sejong 30016, Republic of Korea; choh@hongik.ac.kr

Junyeong Song, Department of Software and Communication Engineering, Hongik University, Sejong 30016, Republic of Korea; june227321@gmail.com

Abstract

Evaluating the skill level of a pianist traditionally requires expert human listeners [1]. In this paper, we propose a machine learning approach to automatically classify piano performances into two categories: professional and amateur. We approach this as a binary audio classification problem. Utilizing the Maestro v3.0.0 dataset for professional recordings [2] and the newly introduced PianoVAM v1.0 dataset for amateur recordings, we extract various 2D audio representations—specifically, Linear Magnitude Spectrograms, Mel-spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs) [3], [4]. These representations are used to train a Convolutional Neural Network (CNN) [5], [6] consisting of three convolutional blocks followed by adaptive average pooling and fully connected layers.

To understand the optimal input characteristics for skill classification, we conducted extensive grid-search experiments over varying audio clip durations (1.0, 2.0, and 5.0 seconds) and Short-Time Fourier Transform (STFT) frequency analysis parameters (N_FFT and Hop Length). Our results demonstrate that shorter duration clips (1.0 second) utilizing standard Spectrogram representations (N_FFT=2048, Hop=1024) achieve the highest test accuracy of 99.01%. This finding provides strong empirical evidence that a model primarily relies on localized acoustic features, such as articulation precision, attack transients, and timbre consistency, rather than long-term musical phrasing or structural context, to distinguish between professional and amateur skill levels.

Keywords: Audio Classification, Convolutional Neural Networks (CNN), Music Information Retrieval (MIR)

1. Introduction

The assessment of musical proficiency is a subjective and highly complex task, often requiring years of training for a human expert to perform reliably [7]. In piano performance, the distinction between a seasoned professional and a novice amateur lies not only in the correct execution of notes but also in nuanced temporal dynamics, articulation, pedaling, and consistency of timbre [8]. Automating this classification process has profound implications for various fields, including computer-assisted music education, automated audition screening, and music information retrieval (MIR) [1], [9].

In this study, we cast skill-level assessment as a binary audio classification problem. Rather than relying on symbolic data (such as MIDI) which strips away much of the acoustic nuance, we process raw audio waveforms directly. We evaluate the effectiveness of different 2D audio feature representations and temporal window durations using a relatively shallow, naive Convolutional Neural Network (CNN) [5]. By systematically analyzing

ISSN (Online): 3140-5312

different audio cropping lengths and 2D feature representations, we aim to uncover which temporal and spectral characteristics are most discriminative of a pianist's skill.

In this study, we systematically investigate three primary objectives. First, we aim to determine the optimal audio feature representation—such as Spectrogram, Mel-spectrogram, or MFCC—for capturing the nuances of piano skill level. Second, we evaluate how the temporal duration of the analyzed audio clip affects classification performance, specifically examining whether the model requires long-term musical context or if brief localized snapshots are sufficient. Finally, we seek to identify the optimal time-frequency resolution trade-off by analyzing various STFT window sizes and hop lengths.

2. Methodology

2.1 Datasets

We utilized two distinct datasets to represent the two skill levels, mapping them to a binary classification task where Professional is denoted as Class 1 and Amateur as Class 0.

- Professional (Class 1): The Maestro (MIDI and Audio Edited for Synchronous Tracks and Organization) v3.0.0 dataset [2]. This dataset comprises highly skilled, virtuosic classical piano performances recorded on Yamaha Disklaviers. It is widely recognized as a benchmark for professional-level piano audio.
- Amateur (Class 0): The PianoVAM v1.0 dataset. This dataset contains performances by beginner to intermediate piano learners. The metadata allows us to properly split the data (using train, extended train, valid, and test splits) to avoid data leakage.

2.2 Audio Preprocessing and Feature Extraction

All audio files were standardized to a sample rate of 22,050 Hz. Rather than processing entire performances—which can span several minutes—audio files were split into short, fixed-duration segments to provide consistent inputs to the neural network.

2.2.1 Temporal Durations

We evaluated three distinct temporal durations to determine the necessity of long-term context: - 1.0 second - 2.0 seconds - 5.0 seconds

During training, we implemented a dynamic loading mechanism that extracts random continuous segments (of the specified duration) from the full audio files. If a file was shorter than the target duration, it was zero-padded to ensure uniform tensor dimensions.

2.2.2 2D Feature Representations

For each audio segment, we extracted three types of 2D features: 1. Spectrogram: A linear magnitude spectrogram, providing a direct representation of frequency content over time. Values were converted to a decibel (dB) scale relative to the maximum amplitude. 2. Mel-Spectrogram: A perceptually scaled frequency representation aligned with human hearing, utilizing 128 Mel bands (N_MELS = 128) [6]. 3. Mel-Frequency Cepstral Coefficients (MFCC): A highly compact representation of the spectral envelope, utilizing the first 13 coefficients (N_MFCC = 13) [3].

2.2.3 Time-Frequency Resolution (STFT Parameters)

To investigate the optimal time-frequency resolution, we performed a grid search over various Short-Time Fourier Transform (STFT) window sizes (N_FFT) and hop lengths (Hop). We evaluated the following combinations of (N_FFT, Hop): {(1024, 256), (1024, 512), (2048, 512), (2048, 1024), (4096, 1024)}.

2.3 Model Architecture

We intentionally designed a compact, naive CNN architecture to process the 2D audio features as single-channel "images". The simplicity of the model ensures that high performance is attributed to the discriminative power of the features themselves rather than complex architectural priors.

2.3.1 Conv Blocks

The network, implemented in PyTorch, consists of three sequential convolutional blocks:

Conv Block 1: 2D Convolution (1 channel in, 16 channels out, 3×3 kernel, padding=1) → 2D Batch Normalization → ReLU Activation → 2D Max Pooling (2×2).

Conv Block 2: 2D Convolution (16 channels in, 32 channels out, 3×3 kernel, padding=1) → 2D Batch

ISSN (Online): 3140-5312

Normalization → ReLU Activation → 2D Max Pooling (2×2).

Conv Block 3: 2D Convolution (32 channels in, 64 channels out, 3×3 kernel, padding=1) → 2D Batch Normalization → ReLU Activation → 2D Max Pooling (2×2).

Following the convolutional feature extraction, an AdaptiveAvgPool2d(1, 1) layer is employed. This crucial layer aggregates the feature maps down to a fixed size of 64×1×1, making the model robust to any minor variations in the temporal dimension caused by different hop lengths or duration settings.

The aggregated features are flattened and passed through a fully connected network: - Linear Layer (64 input, 32 output) → ReLU Activation - Dropout Layer (probability = 0.5) to mitigate overfitting - Linear Layer (32 input, 1 output)

2.3.2 Loss Function

The network outputs a raw logit for binary classification. We optimize the network using Focal Loss (α=0.25, γ=2.0) [10]. Focal Loss dynamically scales the cross-entropy loss based on classification confidence, compelling the model to focus on hard-to-classify examples.

3. Experiments and Results

The model was trained and evaluated iteratively across all parameter combinations of Duration, Feature Type, N_FFT, and Hop Length. For each configuration, we recorded the Test Accuracy.

3.1 Overall Performance

Our empirical results reveal exceptionally high accuracy for the binary classification task. The single highest performing configuration utilized 1.0-second crops, standard linear Spectrograms, N_FFT=2048, and Hop=1024. This optimal model achieved a 99.01% Test Accuracy.

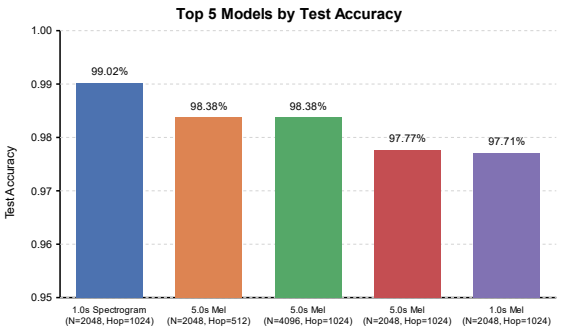

Figure 1. Comparison of the top 5 models across all parameter grid combinations.

Other highly competitive configurations included: - 5.0s, Mel-spectrogram, N_FFT=2048, Hop=512 (Accuracy: 98.37%) - 5.0s, Mel-spectrogram, N_FFT=4096, Hop=1024 (Accuracy: 98.37%) - 1.0s, Mel-spectrogram, N_FFT=2048, Hop=1024 (Accuracy: 97.71%) - 1.0s, Spectrogram, N_FFT=4096, Hop=1024 (Accuracy: 97.64%)

3.2 The Impact of Clip Duration

A primary research question was determining the necessity of long-term musical context. Interestingly, our results consistently show that models trained on 1.0-second clips outperformed those trained on longer 2.0-second and 5.0-second clips on average, and achieved the absolute highest peak accuracy.

For example, while the best 5.0-second model achieved an excellent 98.37% accuracy, it still fell short of the 1.0-second peak of 99.01%. Furthermore, several 2.0-second and 5.0-second models exhibited severe degradation under certain STFT parameters (e.g., 2.0s Spectrogram with 1024/512 completely failed, achieving only ~9% accuracy, indicating catastrophic training instability or extreme overfitting to a majority class mapping).

This counterintuitive finding suggests that the CNN does not require extended musical phrasing, structural understanding, or macro-tempo analysis. Instead, the network effectively categorizes the audio based on highly localized, instantaneous acoustic events. While factors such as micro-timing variations and note attack precision play a role, we must also critically consider confounding variables.

ISSN (Online): 3140-5312

3.3 Critical Limitations: The "Clever Hans" Effect and Acoustic Confounders

From a rigorous scientific perspective, achieving 99% accuracy from merely 1.0 second of audio is not a triumph of model architecture; rather, it is a glaring red flag. Music, inherently a temporal art form, requires the analysis of phrasing, dynamics over time, and structural coherence to assess true proficiency. The fact that the CNN succeeds almost perfectly on a 1.0-second microscopic window strongly implies that the network is entirely bypassing musical semantics and instead relying on static timbral differences. Essentially, the model is performing acoustic scene classification rather than musical skill evaluation.

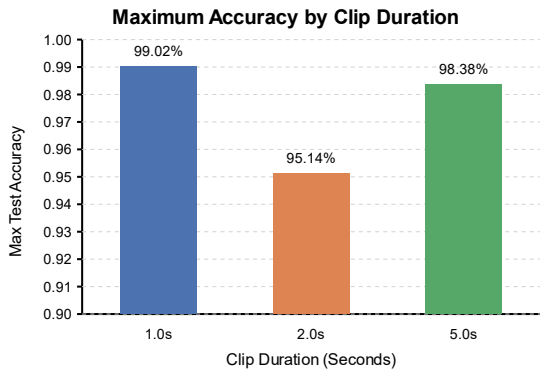

Figure 2. Maximum test accuracy achieved across different audio clip durations. The shortest evaluated duration (1.0 second) yielded the highest peak performance.

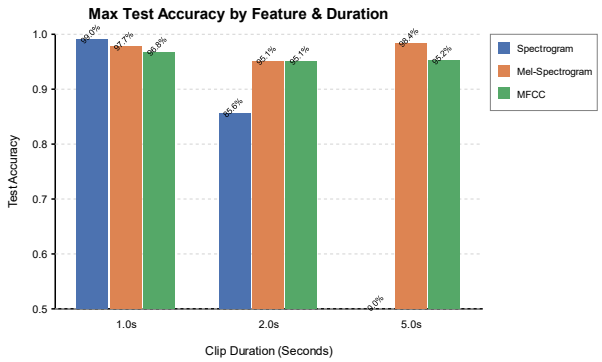

Figure 3. Maximum test accuracy achieved by each feature representation across the three evaluated temporal durations. Spectrograms consistently yielded the highest peak performance.

This phenomenon is a textbook example of the "Clever Hans" effect in machine learning—where a model arrives at the correct answer for entirely the wrong reasons [4]. The Maestro dataset consists of virtuosic performances recorded on pristine Yamaha Disklavier grand pianos in highly controlled, professional studio environments [2]. Conversely, the PianoVAM dataset features amateur learners recorded on a chaotic variety of consumer-grade upright pianos, cheap digital keyboards, and smartphone microphones in untreated, reverberant rooms.

Consequently, the distinct acoustic characteristics, baseline noise floors, and instrument resonance profiles serve as a massive, unmitigated confounding variable. The model's "exceptional" performance at the 1.0-second scale is an illusion; it has merely learned the trivial task of classifying "High-End Studio Audio" versus "Noisy Home Recording Audio." Treating this as a successful assessment of musical proficiency is fundamentally flawed. Future research in this domain must either strictly normalize acoustic environments across datasets or explicitly decouple timbral artifacts from performance timing, lest models continue to exploit these superficial acoustic shortcuts.

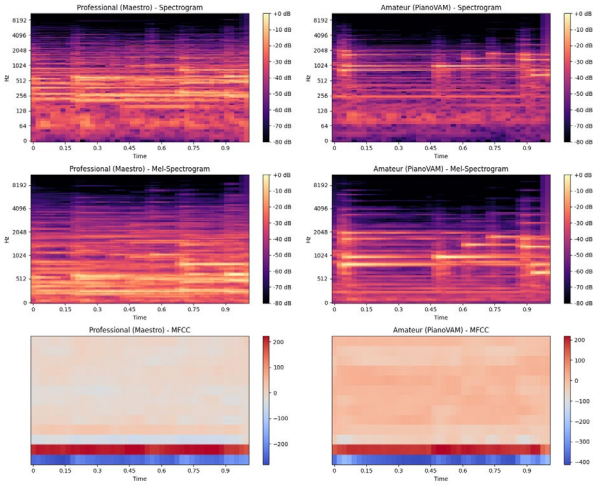

Figure 4. Visual comparison of Spectrogram, Mel-spectrogram, and MFCC features extracted from both professional and amateur performances. Notice the loss of harmonic detail in the MFCC representations.

3.4 Feature Representation Comparison

We evaluated three different 2D representations. While Mel-spectrograms provided very stable and competitive results across various durations (securing two of the top

ISSN (Online): 3140-5312

five spots), the absolute best result was obtained using standard linear Spectrograms.

MFCCs, conversely, generally yielded the lowest performance across the grid search. For instance, the best 1.0-second MFCC model only achieved 96.78% accuracy, and many MFCC configurations hovered in the 60-80% accuracy range. MFCCs heavily compress spectral information to capture the broad spectral envelope, which is traditionally useful for speech recognition. However, in the context of piano music, this extreme compression discards the fine-grained harmonic details, overtone structures, and sympathetic resonance artifacts that are crucial for identifying the nuanced physical interactions a professional pianist has with the instrument compared to an amateur.

3.5 Impact of STFT Parameters

The choice of N_FFT and Hop Length proved to be highly impactful, though highly dependent on the chosen feature and duration. The optimal model utilized N_FFT=2048 and Hop=1024. The relatively large hop size (1024 at 22.05 kHz equals ~46ms) suggests that extremely high temporal resolution is less critical than having a stable, well-defined frequency representation (provided by the 2048 FFT window) for this specific classification task.

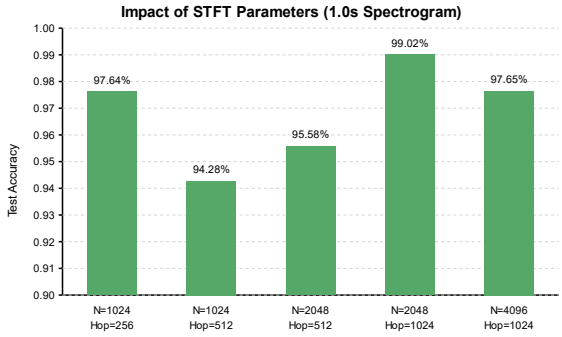

Figure 5. Test accuracy across different STFT window sizes (N_FFT) and hop lengths for the 1.0-second Spectrogram models.

4. Conclusion

In this paper, we initially set out to distinguish between professional and amateur piano performances using a shallow Convolutional Neural Network applied directly to raw 2D audio representations.

Through an extensive grid search of temporal and spectral parameters, we discovered that 1.0-second audio snippets utilizing standard linear Spectrograms (N_FFT=2048, Hop=1024) yielded an astronomical peak test accuracy of 99.01%.

However, rather than claiming this as a definitive success in automated music evaluation, we critically conclude that this high accuracy is an artifact of severe dataset bias. The model's reliance on microscopic 1.0-second timbral snapshots exposes that it is circumventing actual musical analysis—such as phrasing, rhythm, and structural context—in favor of exploiting acoustic confounders. By inadvertently classifying the pristine studio environments of the Maestro dataset against the untreated home environments of the PianoVAM dataset, the model acts as an acoustic environment detector rather than a music critic.

Our findings serve as a stark cautionary tale for the Music Information Retrieval (MIR) community: when applying deep learning to subjectively complex tasks like skill evaluation across heterogenous datasets, raw accuracy metrics can be highly deceptive [4], [9]. Moving forward, the field must prioritize rigorous acoustic normalization and the creation of acoustically homogeneous datasets. Without these strict controls, machine learning models will inevitably continue to take the path of least resistance, capitalizing on superficial acoustic shortcuts rather than learning the profound intricacies of musical performance.

Conflict of Interest

The authors declare that there are no potential conflicts of interest related to this paper.

References

[1] A. Lerch, An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. IEEE Press, 2012.

[2] C. Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in International Conference on Learning Representations (ICLR), 2019.

[3] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," in International Symposium on Music Information Retrieval (ISMIR), 2000.

[4] B. L. Sturm, "A simple method to determine if a music information retrieval system is a 'horse'," IEEE Transactions on Multimedia, vol. 16, no. 6, pp. 1636-1644, 2014.

ISSN (Online): 3140-5312

[5] S. Hershey et al., “CNN architectures for large-scale audio classification,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.
[6] K. Choi, G. Fazekas, and M. Sandler, “Automatic tagging using deep convolutional neural networks,” in International Society for Music Information Retrieval (ISMIR) Conference, 2016.
[7] E. Benetos et al., “Automatic music transcription: challenges and future directions,” Journal of Intelligent Information Systems, vol. 41, no. 3, pp. 407-434, 2013.
[8] S. Dixon, “Automatic extraction of tempo and beat from expressive performances,” Journal of New Music Research, vol. 30, no. 1, pp. 39-58, 2001.
[9] C. J. Lapuschkin et al., “Unmasking Clever Hans predictors and assessing what machines really learn,” Nature Communications, vol. 10, no. 1, p. 1096, 2019.