# HOW DOES THE TEACHER RATE? OBSERVATIONS FROM THE NEUROPIANO DATASET

**Huan Zhang**[1] **Vincent Cheung**[2] **Hayato Nishioka**[2] **Simon Dixon**[1] **Shinichi Furuya**[2]
[1] Queen Mary University of London, Centre for Digital Music
[2]Sony Computer Science Laboratories, Tokyo, Japan

## ABSTRACT

This paper provides a detailed analysis of the NeuroPiano dataset, which comprises 104 audio recordings of student piano performances accompanied with 2255 items of textual feedback and ratings given by professional pianists. We offer a statistical overview of the dataset, focusing on the standardization of annotations and inter-annotator agreement across 12 evaluative questions concerning performance quality. We also explore the predictive relationship between audio features and teacher ratings via machine learning, as well as annotations provided for text analysis of the responses.

## 1. INTRODUCTION

Music Information Retrieval (MIR) has become instrumental in enhancing music education by enabling personalized learning experiences and automating feedback mechanisms [1–4]. Meanwhile, accurately imitating human teachers' feedback [5–7] has been a central goal in MIR-assisted music education, particularly in the context of instrumental performance that involves expressive nuances [8–12]. Moreover, studies have demonstrated the feasibility of using deep-learning based models to assess performance quality objectively and consistently [13–17]. This paper examines the NeuroPiano dataset [5], a collection of student piano performances of technical exercises, annotated with feedback in audio, textual, and rating score modalities.

In this report, we give a statistical overview of the dataset's content regarding audio, text, and score modalities, and explore their relationships. Starting with a detailed examination of rating consistency [18] and distribution, we also annotate the key concepts that manifested in the textual feedback. We also attempt to predict the teacher's rating from audio content. By analyzing how different modalities correlate with teacher assessments, this study contributes to the ongoing discussion about the effectiveness of MIR technologies in educational contexts.

The NeuroPiano dataset [1] , recorded and annotated by the Music Excellence Project at Sony CSL, Tokyo [2] , comprises 104 on-site audio recordings from 39 advanced student pianists performing six standardized technical exercises (including scales, arpeggios, dyads, block chords, octaves) on a Shigeru Kawai grand piano. Each recorded performance is associated with 12 questions addressing multiple performance dimensions from tempo to dynamics to articulation, and annotated by 45 professional pianists. Annotators answered each question by providing a textual response in Japanese, as well as a rating on a 6-point scale.

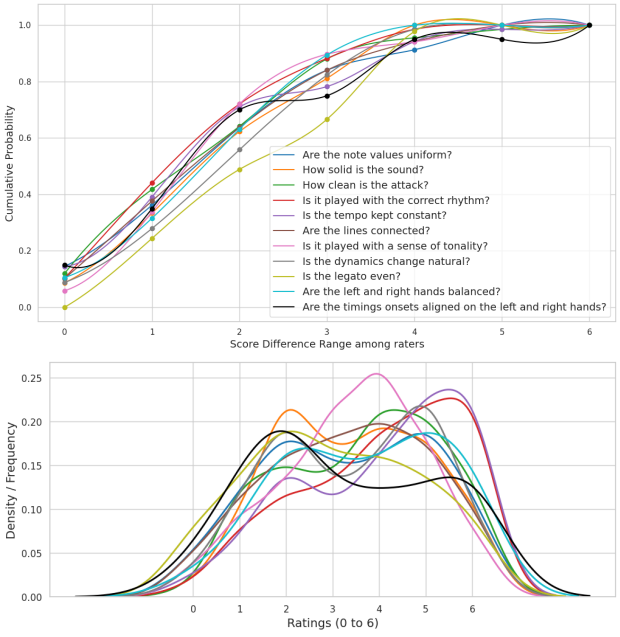


**Figure 1**. Top: Cumulative distribution of the raters' score difference by each question (smoothed); Bottom: Rating distribution (KDE) by each question (sharing same legend )

The dataset includes 2255 audio-question-answer (AQA) triplets, with 391 triplets labeled by one, 874 triplets annotated by two (437 unique), and 990 triplets annotated by three (330 unique) annotators. For those with two or more annotators, we checked for the range between multiple ratings to examine consistency of human judges with this type of question and assessment. For most of the questions, around 65% of the data reached a rough agreement between annotators (with differences in ratings $\leq 2$), although questions on dynamics and legato were the most controversial, with only 50% of the data reaching agreement, and almost

[1] https://huggingface.co/datasets/anusfoil/NeuroPiano-data

[2] https://www.sonycsl.co.jp/tokyo/10996/



Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

| idx | Question | Support | Symbolic feats. MAE | Symbolic feats. model | Audio feats. MAE | Audio feats. model | Dummy-MAE |
|---|---|---|---|---|---|---|---|
| 1 | Are the note values uniform? | 91 | **1.13** | RF (n=100) | 1.27 | RF (n=200) | 1.31 |
| 2 | How solid is the sound? | 90 | 1.14 | SVR (C=0.1) | **1.14** | RF (n=200) | 1.16 |
| 3 | How clean is the attack? | 95 | **1.06** | SVR (C=0.1) | 1.11 | SVR (C=0.1) | 1.21 |
| 4 | Is it played with the correct rhythm? | 94 | **1.16** | SVR (C=1) | 1.21 | SVR (C=0.1) | 1.27 |
| 5 | Is the tempo kept constant? | 88 | **1.23** | SVR (C=0.1) | 1.25 | SVR (C=0.1) | 1.31 |
| 6 | Are the lines connected? | 90 | **1.15** | SVR (C=1) | 1.26 | SVR (C=0.1) | 1.26 |
| 7 | Is it played with a sense of tonality? | 95 | 1.04 | SVR (C=0.1) | **1.02** | SVR (C=0.1) | 1.05 |
| 8 | Is the dynamics change natural? | 89 | **1.27** | SVR (C=0.1) | 1.30 | SVR (C=0.1) | 1.28 |
| 9 | Is the legato even? | 52 | **1.19** | SVR (C=0.1) | 1.21 | SVR (C=0.1) | 1.19 |
| 10 | Are the left and right hands balanced? | 31 | **1.09** | RF (n=200) | 1.36 | SVR (C=0.1) | 1.35 |
| 11 | Are the timing onsets aligned on LH and RH? | 29 | **1.38** | SVR (C=10) | 1.45 | RF (n=100) | 1.57 |

**Table 1**. Best prediction results from symbolic and audio features. Dummy-MAE is the random guessing baseline.

**Q:** Is the legato even?
**A:** The ascending legato is smooth, but the sound breaks at the point where the 1st finger crosses under in the descending passage.

**Figure 2**. An example of concept annotation in teachers' responses. Red: Location; Yellow: Physicality; Green: Technique; Blue: Description.

no data perfectly agreeing between annotators. Figure 1 (top) shows an interpolated CDF plot of the consistency ratings across different questions. The examples that deviate greatly ($\geq 5$) are shown in the demo page.

The distributions of teachers' ratings were bimodal. The KDE plot in Figure 1 bottom shows that most of the questions have peaks in ratings of 2 or 5, indicating a spread between good and bad aspects of performances according to teachers' opinions.

## 2. AUDIO AND RATING

We explore the relationship between audio performance and rating by predicting the rating with various features using regression. To obtain a convincing training target, we first filtered out responses of the same recording whose ratings differed $> 3$ across raters (since that implies a contrast in opinion), and then averaged the remaining ratings. This left us with 844 AQR (audio-question-rating) triplets.

Two types of features were extracted: First, audio features were extracted directly from librosa and madmom, namely MFCC-13, chroma, onset envelope, tempogram, and beat estimation. Second, symbolic features, were obtained by first transcribing the clips [19], and then calculating a set of designed features based on partitura [20] note arrays. Then after grouping the notes on the same onset, we computed the inter- and intra-group onset differences, velocity differences and duration differences, aggregated into mean and standard deviation.

Scipy-implemented regression models, `SVR`, `MLP`, `RandomForest` (`RF`), and `GradientBoosting` (`GB`) were used, with robust scaling, grid search and cross validation as part of the processing pipeline. We also experimented with different feature selections and combinations, where the best results are presented in Table 1. For experimental details please refer to our codebase [3] . The baseline

[3] https://github.com/anusfoil/neuropiano-data-processing


**Figure 3**. Wordcloud for the descriptions in the dataset.

we are comparing against is a dummy regressor that always predicts the mean of the training set.

Our results show that under traditional MIR feature engineering, not all questions could be fitted. Q8 (*dynamics change*) and Q9 (*legato*) did not show any improvement from the dummy baseline, but at the same time they were also the most controversial as shown in Section 1. Likewise, Q2 (*solidity*) and Q7 (*tonality*) only showed a marginal improvement from dummy baseline.

## 3. RATINGS AND TEXTUAL RESPONSES

We next verified the consistency of teachers' ratings and their implied sentiment, by computing a weighted mean of a 5-point BERT sentiment analysis model as in [5]. This yielded a Pearson's correlation of 0.66, for which a box plot of the distribution is shown in our project page [4] .

One of the key applications of our dataset is to reveal how teachers comment on performance. Utilizing `GPT-4o`, we annotated important concepts that were informative within a feedback, namely location, physicality, techniques, and description, as shown in Figure 2. Observations on 'physicality' gave descriptors such as *left hand*, *thumb*, etc, while the 'location' usually involved *descent*, *beginning*, etc.. Figure 3 shows the wordcloud for the prominent descriptions. Future research should leverage this response annotation to gain deeper insights into the feedback process, and we hope the rating, audio and text in the NeuroPiano dataset will help enhance automated analysis of performance critique.

[4] https://river-blackberry-7de.notion.site/NeuroPiano-data-e59586a44d834ca6bd728a5e2b633880?pvs=4

## 4. REFERENCES


[1] H. Zhang, J. Liang, and S. Dixon, "From audio encoders to piano judges: Benchmarking performance understanding for solo piano," in *Proceeding of the 25t International Society on Music Information Retrieval (ISMIR)*, 2024.

[2] H. Kim, P. Ramoneda, M. Miron, and X. Serra, "An overview of automatic piano performance assessment within the music education context," *International Conference on Computer Supported Education, CSEDU - Proceedings*, vol. 1, 2022.

[3] V. Eremenko, A. Morsi, J. Narang, and X. Serra, "Performance assessment technologies for the support of musical instrument learning," *Proceedings of the 12th International Conference on Computer Supported Education (CSME)*, 2020.

[4] A. Morsi, K. Tatsumi, A. Maezawa, T. Fujishima, and X. Serra, "Sounds Out of Pläce? Score-independent detection of conspicuous mistakes in piano performances," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[5] H. Zhang, V. Cheung, H. Nishioka, S. Dixon, and S. Furuya, "LLaQo: Towards a query-based coach in expressive music performance assessment," 2024, Arxiv preprint arXiv:2409.08795.

[6] A. Morsi, H. Zhang, A. Maezawa, S. Dixon, and X. Serra, "Simulating piano performance mistakes for music learning," *In Proceedings of the Sound and Music Computing Conference (SMC)*, 2024.

[7] M. Matsubara, R. Kagawa, T. Hirano, and I. Tsuji, "CROCUS: Dataset of musical performance critiques," in *In Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2021.

[8] H. Zhang, S. Chowdhury, C. E. Cancino-Chacón, J. Liang, S. Dixon, and G. Widmer, "DExter: Learning and controlling performance expression with diffusion models," *Applied Sciences*, vol. 14, no. 15, 2024.

[9] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "An interdisciplinary review of music performance analysis," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 221–245, 2020.

[10] H. Zhang and S. Dixon, "Disentangling the Horowitz factor: Learning content and style from expressive piano performance," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.

[11] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer, "Computational models of expressive music performance: A comprehensive and critical review," *Frontiers in Digital Humanities*, vol. 5, no. October, pp. 1–23, 2018.

[12] H. Zhang, J. Tang, S. Rafee, S. Dixon, and G. Fazekas, "ATEPP: A dataset of automatically transcribed expressive piano performance," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 2022.

[13] J. Huang, Y.-N. Hung, A. Pati, S. K. Gururani, and A. Lerch, "Score-informed networks for music performance assessment," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[14] K. A. Pati, S. Gururani, and A. Lerch, "Assessment of student music performances using deep neural networks," *Applied Sciences*, vol. 8, no. 4, 2018.

[15] H. Zhang, Y. Jiang, T. Jiang, and P. Hu, "Learn by referencing: Towards deep metric learning for singing assessment," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[16] X. Jin, W. Zhou, J. Wang, D. Xu, Y. Rong, and S. Cui, "An order-complexity model for aesthetic quality assessment of homophony music performance," 2023, Arxiv preprint arXiv:2301.05908.

[17] P. Parmar, J. Reddy, and B. Morris, "Piano skills assessment," in *IEEE 23th International Workshop on Multimedia Signal Processing (MMSP)*, 2021.

[18] Y. Jiang, "Expert and novice evaluations of piano performances : Criteria for computer-aided feedback," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[19] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[20] C. Cancino-Chacón, S. D. Peter, E. Karystinaios, F. Foscarin, M. Grachten, and G. Widmer, "Partitura: A python package for symbolic music processing," in *Proceedings of the Music Encoding Conference (MEC)*, Halifax, Canada, 2022.