Automatic Evaluation of Aspects of Performance and Scheduling in Playing the Piano

Hila Tamir-Ostrover
Gilad Baruch
Or Peleg
Yonatan Yellin
Maor Rosenberg
hilatamir@gmail.com
giladbaruch@mail.tau.ac.il
orpeleg@mail.tau.ac.il
yellin@mail.tau.ac.il
maor.rosenberg@cs.tau.ac.il
Tel Aviv University
Tel Aviv, Israel

Alexandra Moringen
Kathrin Krieger
Helge Ritter
abarch@techfak.uni-bielefeld.de
Bielefeld University
Bielefeld, Germany

Jason Friedman
jason@tau.ac.il
Tel Aviv University
Tel Aviv, Israel

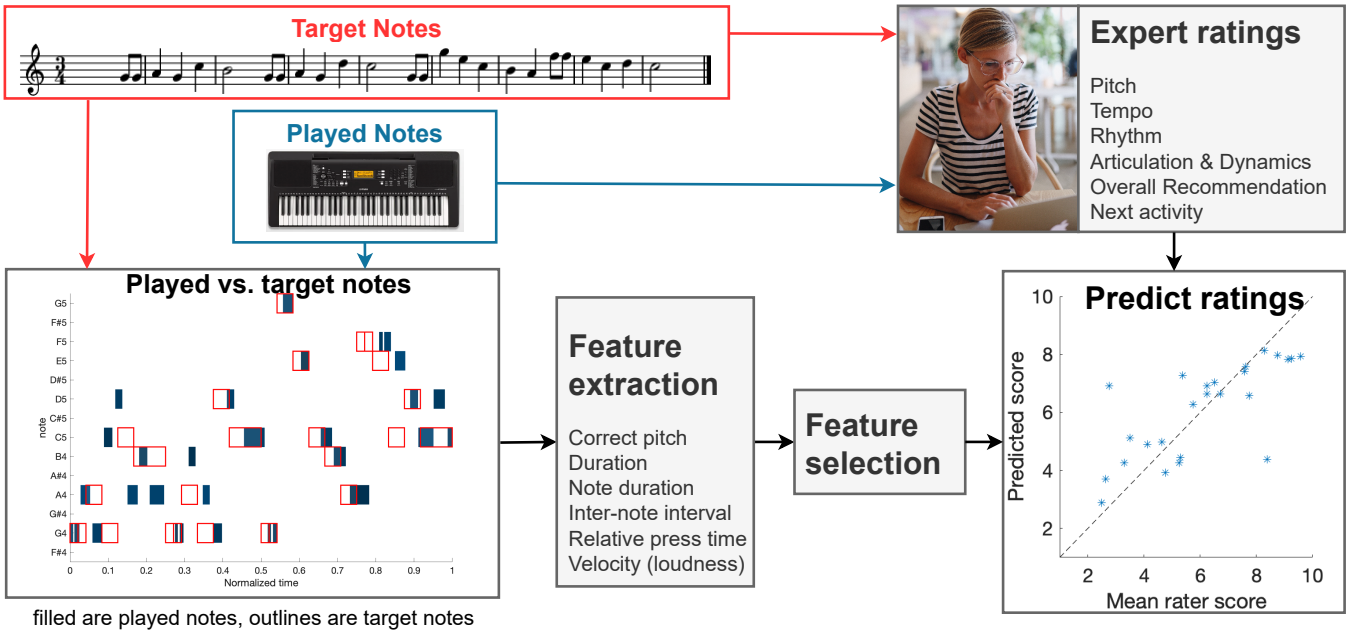

Figure 1: Overview of the process described in the paper: From a set of MIDI recordings, features are extracted based on comparisons to the target notes (musical score). From these features, the most informative are selected, and used to predict ratings of evaluation of different aspects of performance given by experts in the field.

ABSTRACT

There is a growing trend to teach playing an instrument such as a piano at home using an automated system. A key component of such systems is the ability to rate performance of the learner in order to provide feedback and select appropriate exercises. In this study, we expand on previous works that have developed automatic evaluation systems for an overall grade by also providing predictions for specific aspects of performance: pitch, rhythm, tempo, and articulation & dynamics, as well as scheduling what is an appropriate next task. We describe how a set of salient features is extracted

UMAP '22, July 4–7, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9207-5/22/07…$15.00
https://doi.org/10.1145/3503252.3531297

by comparing MIDI performance data of three piano players to an ideal performance, how the features used for evaluation are selected, and evaluate using linear regression how well the selected features are able to predict the mean scores given by a group of domain experts (piano teachers). Relatively good R^2 scores (0.54 to 0.68) are achieved using a small number of features (2 - 4). Such automatic evaluation of different aspects of performance can be used as a part of an automatic learning system, and to help provide learners with detailed feedback on their performance.

CCS CONCEPTS

• **Human-centered computing → Sound-based input / output**; **User models**.

KEYWORDS

piano, evaluation, rhythm, pitch, tempo, dynamics, regression

ACM Reference Format:

Hila Tamir-Ostrover, Gilad Baruch, Or Peleg, Yonatan Yellin, Maor Rosenberg, Alexandra Moringen, Kathrin Krieger, Helge Ritter, and Jason Friedman. 2022. Automatic Evaluation of Aspects of Performance and Scheduling in Playing the Piano. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22), July 4–7, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503252.3531297

1 INTRODUCTION

Computer-aided learning of the piano is becoming more and more popular, largely due to an increasing use of smartphone applications. A necessary feature of such systems is the ability to evaluate performance to provide the learner with appropriate exercises at an appropriate pace. However, it is not obvious how the errors should be computed, or which features of performance should be considered. Rather than devising ad-hoc methods for calculating errors, in this study, we learn how to grade performance based on learning how piano teachers grade performances. Specifically, we examine different aspects of performance, and determine which parameters extracted from performance data are best able to predict the teachers' ratings.

There have been previous efforts to build automatic classifiers of piano and other music performances. In all these studies, expert raters are used to provide ratings for the works - these ratings serve as a test for how well the models perform. A number of studies have used audio data. In a series of studies [19, 27, 30] looking at performances of bands, spectral and rhythmic features were extracted from the audio and support vector regression and deep neural networks were used to predict the grades, with R^2 values compared to experts raters around 0.3-0.5. Other studies using databases of piano recordings from competitions [17, 28], used linear regressions on a number of features extracted from the audio using a convolutional neural network (CNN) to predict the score given by expert raters.

In [20], performances were manually classified into "good", "normal" or "bad" levels, and using a convolutional neural network on microphone recordings of performance, classification accuracy of approximately 90% was achieved. In another study, the skill level of piano players was estimated from YouTube videos of playing the piano which were manually graded by an expert, giving a score of 1 to 10 [18].

Other studies used MIDI recordings (i.e., including key press times, notes played, duration, and key press velocity) thus avoiding the need to extract the performance from noisy audio recordings, and compared the MIDI recordings to the piece they were instructed to play. For example, these techniques have been used to evaluate scales [2, 11] or a particular piece [14], by defining a set of features comparing ideal performance to the played piece, and then using k-nearest neighbors (k-NN) to classify the performance (based on training from expert pianists).

This study is most similar to this last set of studies, but rather than focusing on an overall score, we are interested in evaluating different aspects of performance, including tempo, rhythm, pitch and articulation & dynamics. This is based on rating rubrics used by schools of music [15]. As well as being in accordance with common rating practice at music schools, rating individual aspects of performance will allow us to personalize training for the aspects which require more work, as it is likely that improvement in different facets of performance occurs at different timescales and may vary between people depending on their individual talents and experience. Furthermore, teachers often use teaching and training methods that concentrate on a particular aspect / difficulty of the piece. This holds true not only for beginners (e.g., [25]) but also for advanced students, an early example of which can be found in Cortot's recommended piano exercises for learning to play Chopin [4]. A focus on one of more aspects of the piece to target a particular difficulty has been termed *practice modes* in [10]. Simple online adaptation of the learning content to the specific problems of the learner is becoming common in most learning apps, such as duolingo [5], yousician [31], or piano academy [1]. For example, in duolingo, a task performed with mistakes is given again at the end of the practice episode. In music learning apps such as yousician, the tune is paused until the correct key is hit. In other apps such as lumi [22], a small set of practice modes, such as "slower", or "pitch practice only" exists, but they are selected by the learners themselves, and not based on the performance evaluation with a goal to target a particular type of mistake. Other types of simplification include choosing only specific notes that the learner needs to play, where the rest of the audio content/accompaniment is added by the app. However, this is prescribed for each level, and not calculated depending on the performance/ learning rate of the learner.

In this study, we present a number of features that can be extracted from the MIDI data, and test how well they can predict the scores given by the teachers. These expert ratings are treated as the gold standard because no other assessment is available. Such scores may potentially enable us to schedule different types of practice modes, e.g. focusing on rhythm, tempo or pitch practice, depending on the learner performance evaluation. When presented to the learner, detailed expert-like performance feedback may promote the development of self-assessment skills. Self-assessment of performance is an important metacognitive ability for better practice efficacy. Research suggests that while this ability is more developed for skilled performers [8], it also has a significant impact on novices' practice efficacy and performance [3]. Detailed performance feedback thus has the potential to substantially expedite the student's

skill acquisition in automated learning environments, especially for beginner learners.

2 METHODS

2.1 Participants

Three players (2 beginners, with one year of keyboard experience aged 9 and 14, one adult with 6 years of keyboard experience) played the pieces used in the experiment. The study was approved by the Tel Aviv University Institutional Review Board and the participants (or their parents) signed an informed consent form.

The expert raters consisted of 8 piano teachers. Teachers had a minimum of 10 years of teaching experience.

2.2 Experimental procedure

The participants played the songs on a Yamaha PSR-E363 keyboard, with the scores shown and the MIDI recorded using custom software. The pieces were selected from a beginner's book [24]. The participants (players) played a total of 25 performances of 10 different songs (not all participants played all songs). The order of the songs was the same for all players (following their order in the book).

2.3 Expert ratings

The expert panel graded the anonymized performances in three sections:

- a score from 0 to 4 for the categories of **pitch, tempo, rhythm, and articulation and dynamics** (see Appendix 1),
- an **overall** evaluation of performance - from 1 (lowest) to 10 (highest),
- **what should be the next task** for the learner - either play the same piece (at the same pace, slower, or faster) or play a different piece (easier, same level, or harder).

The scale of 0 to 4 was selected to match the five textual categories in the rubric described in Appendix 1. The overall scale was selected from 1 to 10 to allow the raters more nuance in their grading.

The raters performed the ratings online using the Qualtrics platform [21], where they were shown the sheet music, could play the midi recordings, and mark their ratings. The raters received the pieces in a random order such that subsequent performances were from different players, although the same order was used for all raters. They were able to listen to the piece as many times as they wanted before providing the scores. The duration of each recording was less than one minute, and a single grade (in each category) was given for each recording.

2.4 Performer features

Six types of features (14 in total) were extracted from the recorded MIDI data to test how well they predict the grades given by the teachers. The best mapping between the notes played in the ideal performance and the actual performance was performed using dynamic time warping [13]. This allowed us to know which notes in the ideal performance correspond to which notes from the actual performance. The features were then calculated based on note-wise comparisons.

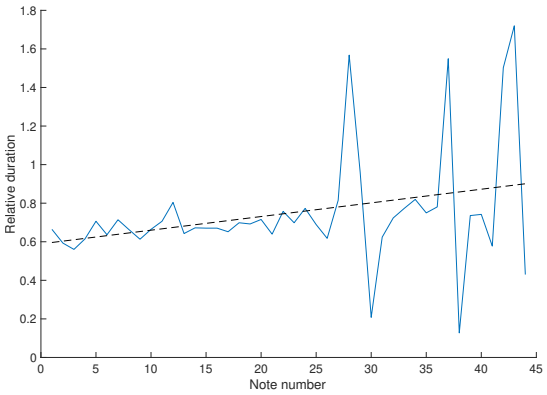

Figure 2: An example of the mean, slope and standard deviation measures extracted from note duration. The graph shows the note duration (normalized to the total time of the piece), such that 1 would indicate the duration was the same as the ideal performance. The dashed line is the regression line. In this example, the mean is 0.749, indicating that on average the duration of the notes was approximately 75% of that of the ideal performance, the slope is 0.0071, indicating that the relative note duration increased over the performance, and the standard deviation (between the regression line and the values) is 0.2880, quantifying the variability in duration produced by the participant (in a perfect performance according to this rating system this value would be zero).

For two of the features (correct pitch and duration), we calculated for each a single value representing the whole performed piece (e.g. proportion of notes played correctly). For the other features (note duration, inter-note intervals, relative press time and loudness), inspired by previous studies [14], we compared the played piece and the ideal piece and calculated three measures - mean, slope and standard deviation to capture how the performance changed throughout the piece, see Figure 2 for an example. The mean measures the mean difference in the quantity over the whole performance, the slope measures how it changes throughout the piece (i.e. it would be zero if it does not change within the song), and the standard deviation measures how performance varies within the song.

2.4.1 Correct pitch. The correct pitch feature was defined as the proportion of notes in the ideal score that were played correctly (considering only the key pressed, and not any other details). This measure may be intuitively understood as an estimate of the accuracy of the performed melodic contour, and could also be understood as the accuracy of the spatial finger location, from a movement-related perspective.

2.4.2 Duration. The duration feature was defined as the relative difference in duration between the played piece and the duration based on the tempo suggested in the songbook, normalized by the ideal duration.

$$\text{Feature}_{\text{duration}} = (\text{Duration}_{\text{actual}} - \text{Duration}_{\text{ideal}})/\text{Duration}_{\text{ideal}}$$

As overall duration is closely related to tempo, this feature may be understood as a rough estimate for performance tempo, relative to the tempo of the ideal performance.

2.4.3 Note duration. The note duration feature was defined as the relative difference in duration between the played notes and ideal notes, normalized to the overall duration This measure roughly reflects how precise the duration of a given note is within the context of the performed tempo. Across a piece, it is capable of giving a note-sensitive estimate of how well a beat was kept (Slope), how appropriate the performance tempo was (Mean), as well as how consistent the rhythms were (SD).

2.4.4 Inter-note intervals. The inter-note interval was defined as the relative difference in onset time between correct consecutive notes played between the played and ideal pieces, both normalized to the appropriate overall notes. Similar to note duration, this feature also measures a given note's timing, only it disregards the note's articulation. Articulation (e.g., *staccato, legato, tenuto*) is the musical parameter that describes the duration of a depressed note with regard to the duration of the inter-onset interval between that note and the following one.

2.4.5 Relative press time. The relative press time measure was defined as the note duration divided by the inter-note interval (from that note to the next). The measure was not calculated for the last note as there is no next note. This feature corresponds with the articulation of the performed notes. A small relative press time reflects shorter articulation, such as *staccato*. A high relative press time will correspond with longer articulations, such as *legato*. It may also highlight "lingering" notes, that are not released on time (in which case the relative press time would be substantially larger than 1).

2.4.6 Velocity (loudness). MIDI keyboards typically record "velocity" - how fast the key is pressed, and this is used to control the loudness of the particular note. This measure is based on the relative difference in velocity between the played and ideal pieces

Note that this study's notated exercises - like many beginner level exercises - did not contain any instructions for dynamics or articulation. Therefore, their associated features' mean or any other comparison to the "ideal performance" is of no relevance in our case. Nevertheless, it may prove relevant to the analysis of more advanced levels of performance. The stability and variation of velocity and articulation, estimated by the standard deviation and slope of their related features, would still be of interest even in the absence of anchoring instructions in the notated exercises.

2.5 Statistical comparisons

We tested the between-rater consistency of the eight raters using Krippendorff's Alpha [9]. To determine which of the proposed features should be used for predicting the ratings (i.e., non-redundant predictors), we constructed the lasso fit [7] using 4-fold cross-validation, and selected those which correspond to the minimum cross-validated mean squared error. Using these selected features, we performed linear regression to find the relevant weights, then looked at the R^2 measure to determine the goodness of fit. We tested for the normality of the residuals using Q-Q plots, and for the absence of multicollinearity using the Variance Inflation Factor (VIF) [6].

The software was implemented using Matlab (Mathworks, version 2021b) and is available online [1]

3 RESULTS

The main goal of the experimental evaluation is to explore which features extracted from the learner performance, such as correct pitch, or inter-note intervals, predict the expert rating the best. However, we first briefly present the characteristics of the raters and the rating categories.

3.1 Between-rater consistency

The between-rater consistency, as quantified using Krippendorff's alpha coefficient, varied between the different features that were graded, see Table 1. While there was relatively good agreement for pitch and overall (0.85 and 0.73), articulation & dynamics showed a relatively low agreement (0.20) between raters.

Pitch	Tempo	Rhythm	Articulation & Dynamics	Overall
0.85	0.62	0.45	0.20	0.73

Table 1: Krippendorff's alpha coefficient for assessing inter-rater agreement, 1 indicates perfect agreement, 0 is chance agreement between raters.

3.2 Correlation between rating categories

There was a high degree of correlation between the ratings for the different categories, see Figure 3. The least amount of correlation (0.77) was found between rhythm and pitch.

3.3 Feature distribution

Histograms of the 14 features calculated are shown in Figure 4.

3.4 Selected features

Table 2 presents the results from the lasso process to identify non-redundant predictors. The process found that *Correct pitch* was predictive for all ratings. *Duration* was predictive for all ratings apart from pitch. The standard deviation of inter-note interval was predictive for all ratings apart from rhythm. The mean of relative note time was predictive for tempo, rhythm and overall. Many of the features were not predictive for any of the ratings, in particular, no velocity-related features were predictive (including for predicting articulation and dynamics).

3.5 Prediction results

The residuals from the linear regression were found to be close to normally distributed using Q-Q plots, and the values for the Variance inflation factor (VIF) were all less than 2.2, suggesting that multicollinearity does not pose a serious problem for the models used here [12].

The predictions for the five ratings are shown in Figure 5, with R^2 values ranging from 0.54 (rhythm) up to 0.68 (pitch), with all p

[1] https://github.com/JasonFriedman/AutomaticPianoEvaluation/

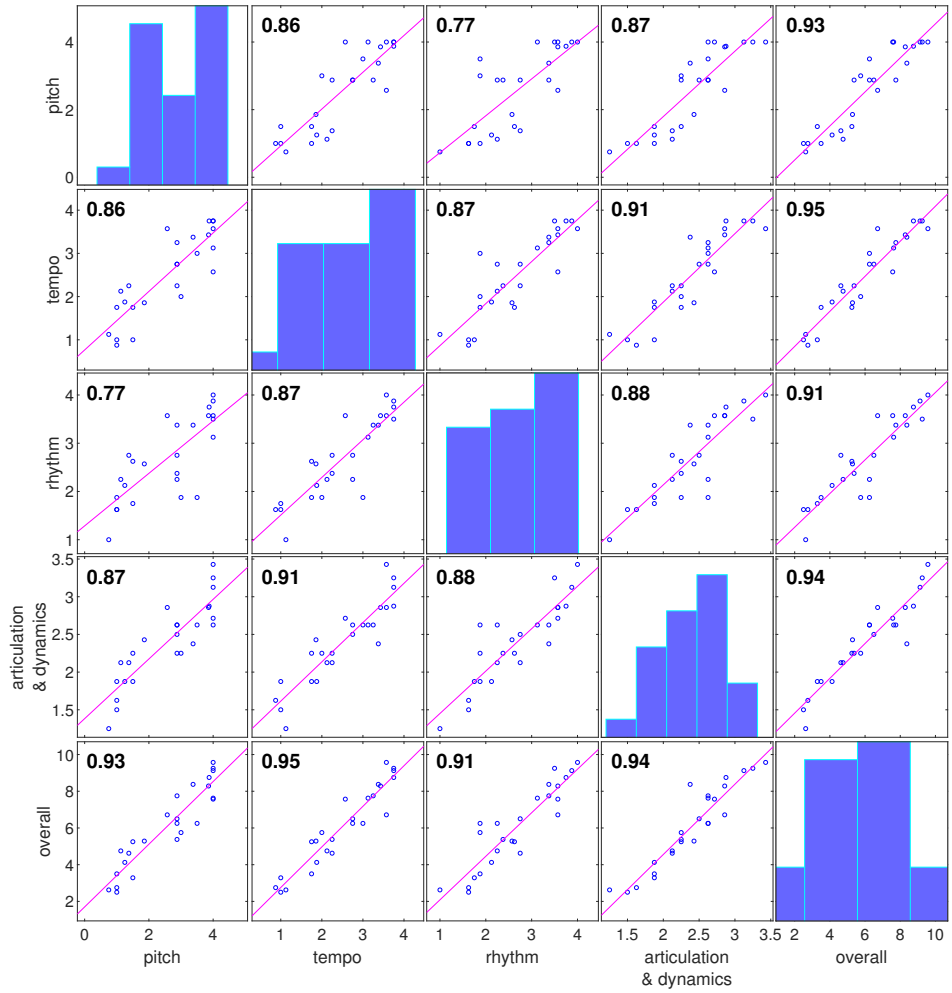

Figure 3: Correlations between the mean ratings for the different categories. The values on the diagonal show the histogram of the relevant quantity (i.e., the relative frequencies, note that the scale of the histograms are normalized). The numbers in each box indicate the Pearson correlation coefficient.

values ≤ 0.01. In addition, the probability of selection of the same or different next piece is shown, as well as the predictions for what to do in the next piece. The regression equations are given in Table 3.

4 DISCUSSION

In this study, we described a set of features that can be extracted from piano MIDI data in order to predict the ratings provided by experts in the field (piano teachers). We found that a relatively small number of features (2 to 4) are sufficient to achieve goodness of fit (R^2) values in the range of 0.54 to 0.68.

The consistency between raters varied among the different categories used. While the ratings were relatively consistent for pitch (Krippendorff's alpha of 0.85) and overall (0.73), the agreement was much lower for articulation and dynamics (0.20). This suggests that this measure may not have been well defined in the rubric, or is difficult to judge based on a MIDI recording - a video recording, for example, might give more information about articulation. Another possibility is that this measure is less relevant for beginner players, and becomes relevant only at a later stage, an idea supported also by the scarcity of articulation and dynamic signs in beginner repertory, which are at times completely omitted ([25] - we discuss this in more detail at the conclusions section). We note that none of the velocity-related features were selected for articulation and dynamics, despite the fact that dynamics (including control of the volume) was part of the description.

The goodness of fit (R^2) values achieved here were similar to those found in previous studies [14, 19], ranging from 0.54 to 0.68. Given the significant amount of variation in the ratings given (covering nearly the whole range for most of the categories), this suggests that linear regression is able to provide a reasonable estimation of performance in the different categories. It is likely that the goodness of fit could be improved by the use of more advanced machine learning regression techniques [23] if more training data was available.

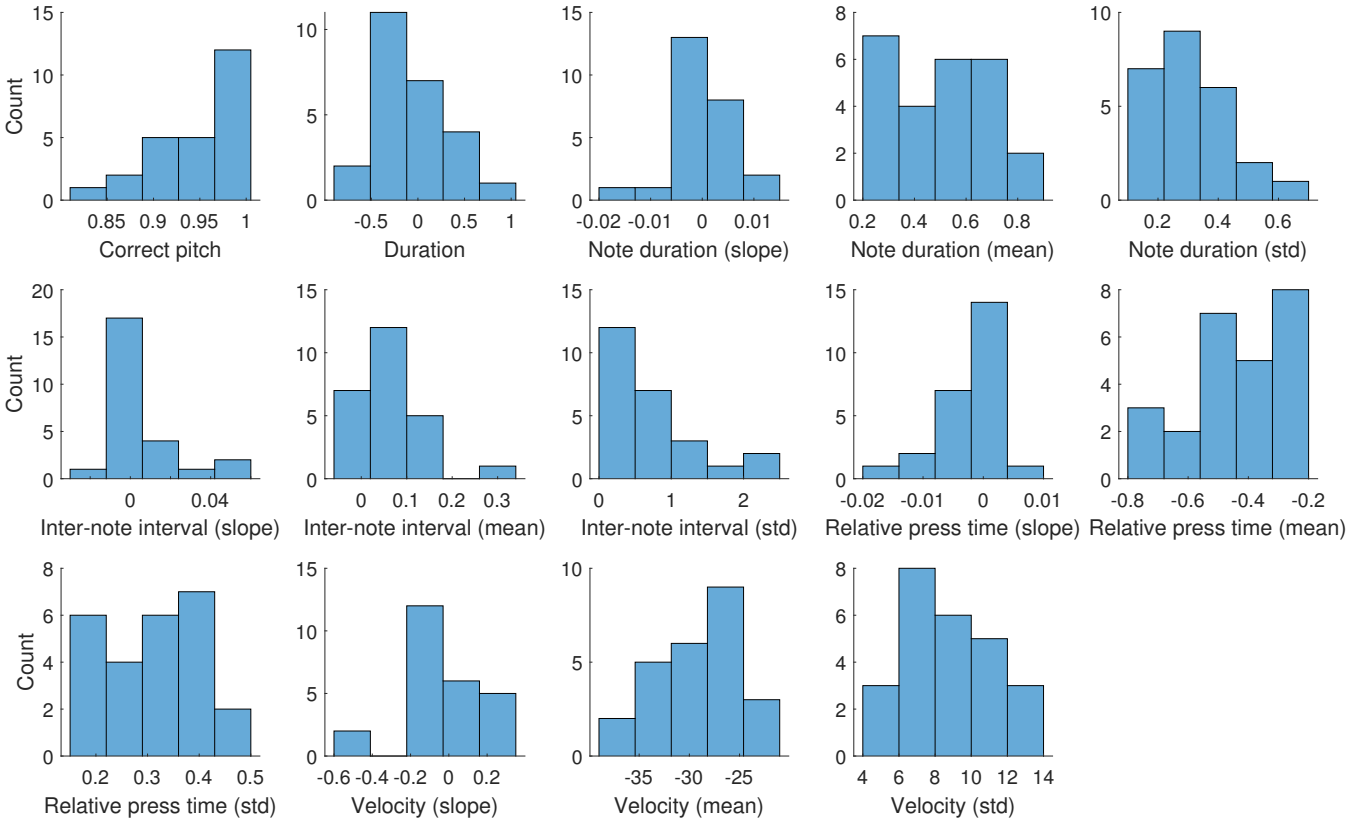

Figure 4: Histograms of the 14 features calculated from the data. Each histogram shows the counts (from a total of 25 performances).

Feature name	Pitch	Tempo	Rhythm	Articulation & Dynamics	Overall	Recommendation	Same choice	Different choice
Correct pitch	X	X	X	X	X	X	X	X
Duration		X	X	X	X			
Note duration (slope)								
Note duration (mean)						X		
Note duration (std)								
Inter-note interval (slope)								X
Inter-note interval (mean)								
Inter-note interval (std)	X	X		X	X	X	X	
Relative press time (slope)								
Relative press time (mean)		X	X		X			
Relative press time (std)								
Velocity (slope)								
Velocity (mean)								
Velocity (std)								

Table 2: X indicates that a particular features will be used in the linear regression, based on the outcome of the lasso fit, except for recommendation, which was based on chi squared tests

There was a relatively high correlation between the different rating categories, see Figure 3. There are several potential reasons for this. First, a learner that is good (or bad) in general will likely be good (or bad) both overall and in many of the components. Second, it may have been difficult for the raters not to be influenced by the overall performance when giving grades on individual components. This may explain why the proportion of correct pitch played was a positive predictor for all ratings. This is true even for ratings which should not be affected by this, e.g. rhythm or tempo. Yet another explanation that may account for the correlation is that especially

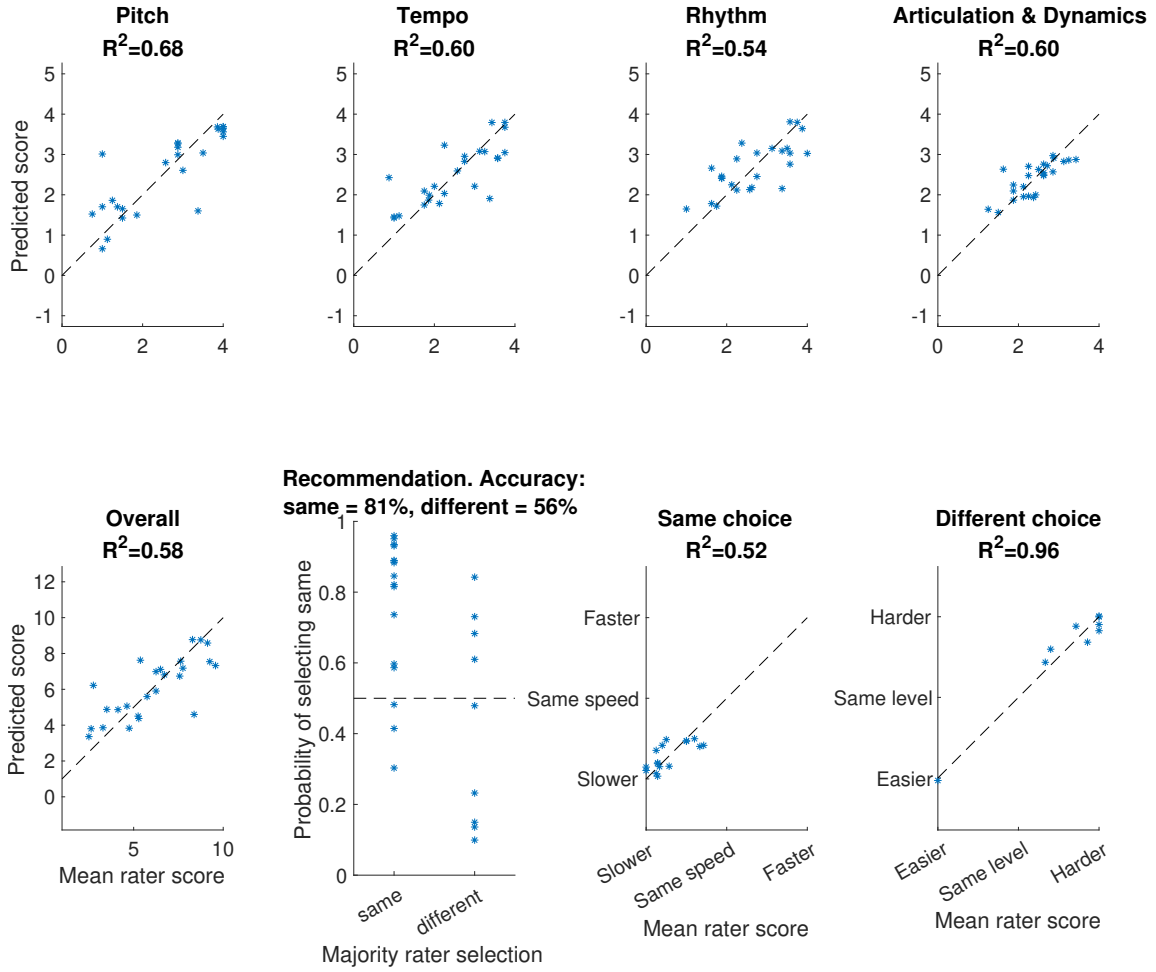

Figure 5: The mean scores of the raters (x-axis) and predicted scores from the linear regression (y-axis)

with beginners, an error in one parameter may cause a cascade of errors in others. Thus for example, playing the wrong pitch may cause a student to pause rhythmically.

This study had a number of limitations. A larger sample size (both in terms of participants and number of pieces played) would be helpful for better determining appropriate predictors, and ideally all performers should play all pieces. The study aimed at checking performance evaluation at a beginner level. Thus, the repertory used and the evaluated MIDI performances were designed as such. Future work should develop and test measures that are appropriate for a more advanced level. One challenge that may require a solution is the adjustment of the measures to multiple simultaneous notes, which is the normal case rather than the exception once a skill level that enables bimanual piano playing is reached. Furthermore, it is possible that performance predictors vary substantially between different skill levels. In particular, expressivity (associated with dynamics and articulation) may play a larger role in the judgment of advanced students' performance. Another implication of this study's focus on beginner repertory is the lack of specific instructions for dynamics, articulations, and tempo. As mentioned above, dynamic and articulations, as well as tempo instructions, are often entirely omitted in beginner repertory (e.g., [25]) or kept constant (e.g., [16, 26]). Yet, no note can be played without a particular dynamic (defined as the velocity or loudness of the pressed note), articulation (defined as the ratio between the duration of the pressed note and the inter-onset interval between a note and its successor), and tempo (beats per minute - BPM). Furthermore, especially for tempo, "proper" values are often implied even in the absence of explicit instructions, based on familiarity with the melodies (which are often well-known nursery or folk songs), or - in the case of unfamiliar melodies - based on their rhythmic profile, meter, and other musical heuristics. Thus, while judging these parameters in beginner level performance is still relevant (as is reflected by the high correlation of tempo rating between experts, for example, in our experiment), it may well be more challenging than in advanced levels, where these instructions are made explicit. Relatedly, in this study, "dynamics and articulation" were presented to teachers and rated as a single category. This is not uncommon in the context of

$$\begin{aligned}
\text{Pitch} &= +14.55 \times \text{Correct pitch} - 0.70 \times \text{Inter-note interval (std)} - 10.81 \\
\text{Tempo} &= +7.62 \times \text{Correct pitch} - 0.68 \times \text{Duration} \\
&\quad - 0.14 \times \text{Inter-note interval (std)} - 1.96 \times \text{Relative press time (mean)} - 5.54 \\
\text{Rhythm} &= +6.75 \times \text{Correct pitch} - 0.74 \times \text{Duration} \\
&\quad - 1.62 \times \text{Relative press time (mean)} - 4.47 \\
\text{Articulation \& Dynamics} &= +5.43 \times \text{Correct pitch} - 0.37 \times \text{Duration} \\
&\quad - 0.16 \times \text{Inter-note interval (std)} - 2.67 \\
\text{Overall} &= +21.71 \times \text{Correct pitch} - 1.11 \times \text{Duration} \\
&\quad - 0.47 \times \text{Inter-note interval (std)} - 3.19 \times \text{Relative press time (mean)} - 15.70 \\
P(same) &= 1/\left(1 + e^{-(8.81 + -13.11\text{Correct pitch} + 7.33\text{Note duration (mean)} + 1.06\text{Inter-note interval (std)})}\right) \\
\text{Same choice} &= +2.41 \times \text{Correct pitch} - 0.18 \times \text{Inter-note interval (std)} - 0.84 \\
\text{Different choice} &= +10.04 \times \text{Correct pitch} + 25.69 \times \text{Inter-note interval (slope)} - 7.19
\end{aligned}$$

Table 3: The regression equations calculated from the data

beginner ratings, due to the assumed secondary importance of these parameters compared to pitch and rhythmic accuracy. Followup studies may consider splitting this rating into two independent ratings, which may provide a more detailed picture of their independent weight in overall performance assessment, and may be more relevant for the evaluation of advanced-level performances.

In addition, the scores were calculated for the piece as a whole, whereas performance may vary across sections of the piece. While in this case, the performances were short (less than 1 minute), for longer performances it would make sense to evaluate the performance of different parts of the performance independently.

4.1 Conclusions

This study demonstrated the feasibility of building an automated rating system for piano learners, whose rating is predictive of experts' evaluations of learners' performances. The system breaks down the performance assessment into several musical subcategories, including pitch, rhythm, tempo, and dynamics and articulation. The reliance on a rubric of parameters detailing different aspects of the overall performance is in accordance with common approaches to music performance evaluation (e.g., [29]). The categories included in our rating represent basic musical dimensions that are both quantitatively measurable and musically relevant for a wide range of proficiency levels. The ability of providing detailed category-based feedback may allow a more efficient design of automated learning curriculum and faster advancement toward skill mastery. Furthermore, when given as feedback to the learner, detailed category-based assessment may facilitate the development of the learner's self-assessment ability - a metacognitive skill enabling efficient practice and improved performance.

ACKNOWLEDGMENTS

This study was funded by the German-Israeli Foundation for Scientific Research and Development (GIF).

REFERENCES

[1] Piano Academy. 2022. Piano Academy. https://www.pianoacademy.app/

[2] Seiko Akinaga, Masanobu Miura, Norio Emura, and Masuzo Yanagida. 2006. Toward realizing automatic evaluation of playing scales on the piano. In *9th International Conference on Music Perception and Cognition*. Bononia University Press, Bologna, Italy, 1843–1847.

[3] Meghan Bathgate, Judith Sims-Knight, and Christian Schunn. 2012. Thoughts on Thinking: Engaging Novice Music Students in Metacognition. *Applied Cognitive Psychology* 26, 3 (2012), 403–409. https://doi.org/10.1002/acp.1842

[4] Frederic Chopin and Alfred Cortot. 1986. *Chopin: 12 Studies for Piano, Op. 10*. BMG Ricordi, New York and Paris.

[5] Duolingo. 2022. Duolingo. https://en.duolingo.com/

[6] Andy Field. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. SAGE Publications Ltd, Los Angeles.

[7] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33 (Feb. 2010), 1–22. https://doi.org/10.18637/jss.v033.i01

[8] Susan Hallam. 2001. The development of metacognition in musicians: Implications for education. *British Journal of Music Education* 18, 1 (March 2001), 27–39. https://doi.org/10.1017/S0265051701000122 Publisher: Cambridge University Press.

[9] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1, 1 (2007), 77–89. https://doi.org/10.1080/19312450709336664

[10] Alexandra Moringen, Sören Rüttgers, Luisa Zintgraf, Jason Friedman, and Helge Ritter. 2021. Optimizing piano practice with a utility-based scaffold. arXiv:2106.12937 http://arxiv.org/abs/2106.12937

[11] Shinya Morita, Norio Emura, Masanobu Miura, Seiko Akinaga, and Masuzo Yanagida. 2009. Evaluation of a scale performance on the piano using spline and regression models. In *International symposium on performance science*. AEC, Auckland, New Zealand, 77–82.

[12] Raymond H. Myers. 1990. *Classical and Modern Regression with Applications*. PWS-KENT, Boston, MA.

[13] Meinard Müller. 2007. Dynamic Time Warping. In *Information Retrieval for Music and Motion*. Springer, Berlin, Heidelberg, 69–84. https://doi.org/10.1007/978-3-540-74048-3_4

[14] Asami Nonogaki, Norio Emura, Masanobu Miura, Seiko Akinaga, and Masuzo Yanagida. 2012. Evaluation parameters for proficiency estimation of piano based on tendency of moderate performance. In *12th International Conference on Music Perception and Cognition and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*. School of Music Studies, Aristotle University of Thessaloniki, Thessaloniki, Greece, 728–737.

[15] Associated Board of the Royal Schools of Music (ABRSM). 2021. ABRSM: Graded music exam marking criteria. https://us.abrsm.org/en/our-exams/information-and-regulations/graded-music-exam-marking-criteria/. Accessed: 2021-01-12.

[16] Willard A Palmer, Morton Manus, and Amanda Vick Lethco. 1995. *Alfred's Basic Adult Piano Course Level One: Lesson Book*. Alfred Publishing Company, Van Nuys, California.

[17] Jing Pan, Ming Li, Zhanmei Song, Xin Li, Xiaolin Liu, Hua Yi, and Manman Zhu. 2017. An Audio Based Piano Performance Evaluation Method Using Deep Neural Network Based Acoustic Modeling. In *Interspeech 2017*. ISCA, Stockholm, Sweden, 3088–3092. https://doi.org/10.21437/Interspeech.2017-866

[18] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. 2021. Piano Skills Assessment. arXiv:2101.04884 http://arxiv.org/abs/2101.04884

[19] Kumar Ashis Pati, Siddharth Gururani, and Alexander Lerch. 2018. Assessment of Student Music Performances Using Deep Neural Networks. *Applied Sciences* 8, 4 (April 2018), 507. https://doi.org/10.3390/app8040507 Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[20] Varinya Phanichraksaphong and Wei-Ho Tsai. 2021. Automatic Evaluation of Piano Performances for STEAM Education. *Applied Sciences* 11, 24 (2021), 11783. https://doi.org/10.3390/app112411783 Number: 24 Publisher: Multidisciplinary Digital Publishing Institute.

[21] Qualtrics. 2022. Qualtrics XM. https://qualtrics.com

[22] ROLI. 2022. Let's get started with Lumi and Roli Studio. https://roli.com/lumi-start

[23] Freek Stulp and Olivier Sigaud. 2015. Many regression algorithms, one unified model: A review. *Neural Networks* 69 (Sept. 2015), 60–79. https://doi.org/10.1016/j.neunet.2015.05.005

[24] Yigal Tav-El. 2001. *Organit im kol shir (Organ with every song) - in Hebrew.* Zemer Am, Israel.

[25] John Thompson. 1955. *John Thompson's Easiest Piano Course, Book 1.* Willis Music Company, Florence, Kentucky.

[26] John Thompson. 2009. *John Thompson's Modern Course for the Piano: the First Grade Book : Something New Every Lession.* Willis Music, Florence, Kentucky.

[27] Amruta Vidwans, Siddharth Gururani, Chih-Wei Wu, Vinod Subramanian, Rupak Vignesh Swaminathan, and Alexander Lerch. 2017. Objective descriptors for the assessment of student music performances. In *Conference on Semantic Audio.* Audio Engineering Society, Erlangen, Germany, 9.

[28] Weiqing Wang, Jin Pan, Hua Yi, Zhanmei Song, and Ming Li. 2021. Audio-Based Piano Performance Evaluation for Beginners With Convolutional Neural Network and Attention Mechanism. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1119–1133. https://doi.org/10.1109/TASLP.2021.3061267 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[29] Brian C. Wesolowski. 2012. Understanding and Developing Rubrics for Music Performance Assessment. *Music Educators Journal* 98, 3 (March 2012), 36–42. https://doi.org/10.1177/0027432111432524 Publisher: SAGE Publications Inc.

[30] Chih-Wei Wu, Siddharth Gururani, Christopher Laguna, Ashis Pati, Amruta Vidwans, and Alexander Lerch. 2016. Towards the Objective Assessment of Music Performances. In *International Conference on Music Perception and Cognition.* ICMPC14, San Francisco, CA, 99–102. https://www.annualreviews.org/doi/10.1146/annurev.psych.48.1.115

[31] Yousician. 2022. Yousician. https://yousician.com/

A QUESTIONNAIRE GIVEN TO THE TEACHERS

	4	3	2	1	0
Pitch	Student had no pitch mistakes	Student had a couple of minor (single note) pitch mistakes	Student had more than two small pitch mistakes, or at least one substantial (multiple-note) mistake	Student had many pitch mistakes, but some sequences were still correct	Melody was almost unrecognizable
Tempo	Student performed at or near an appropriate tempo and maintained a steady tempo throughout	Student performed at or near an appropriate tempo and mostly maintained a steady tempo, with one or two fluctuations or a minor drift	Student either performed at a considerably different tempo than appropriate for the piece, stopped playing at one point, or had multiple fluctuations, but was able to maintain tempo between these events	Student did not maintain a consistent tempo most of the time, stopped multiple times, or had many fluctuations	Student could not maintain a steady beat at all, or could not complete the performance of the piece
Rhythm	Student had no rhythmic mistakes	Student had a couple of minor rhythmic inaccuracies or mistakes	Student had more than two minor rhythmic mistakes, or one major rhythmic mistake	Student had many minor rhythmic mistakes or several major rhythmic mistakes	Rhythm was almost unrecognizable
Articulation & Dynamics	Accurately follows Articulation & Dynamics in the score when they are indicated, and shows control of Articulation & Dynamics overall	Mostly controlled use of articulation and dynamics. Close to the score when they are indicated, with some deviations	Partly controlled use of articulation and dynamics: Partly following the score, or showing occasional sensitive manipulation of Articulation & Dynamics but not consistently	Very little use of Articulation & Dynamics. Almost doesn't follow the score, or flat articulation and dynamics	No Control

Table 4: Rating scales used by the expert raters