TOWARDS EXPLAINABLE AND INTERPRETABLE MUSICAL DIFFICULTY ESTIMATION: A PARAMETER-EFFICIENT APPROACH

Pedro Ramoneda[1] **Vsevolod Eremenko**[1] **Alexandre D'Hooge** [2]
Emilia Parada-Cabaleiro[3] **Xavier Serra**[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[3] Department of Music Pedagogy, Nuremberg University of Music, Germany
pedro.ramoneda@upf.edu

ABSTRACT

Estimating music piece difficulty is important for organizing educational music collections. This process could be partially automatized to facilitate the educator's role. Nevertheless, the decisions performed by prevalent deep-learning models are hardly understandable, which may impair the acceptance of such a technology in music education curricula. Our work employs explainable descriptors for difficulty estimation in symbolic music representations. Furthermore, through a novel parameter-efficient white-box model, we outperform previous efforts while delivering interpretable results. These comprehensible outcomes emulate the functionality of a rubric, a tool widely used in music education. Our approach, evaluated in piano repertoire categorized in 9 classes, achieved 41.4% accuracy independently, with a mean squared error (MSE) of 1.7, showing precise difficulty estimation. Through our baseline, we illustrate how building on top of past research can offer alternatives for music difficulty assessment which are explainable and interpretable. With this, we aim to promote a more effective communication between the Music Information Retrieval (MIR) community and the music education one.

1. INTRODUCTION

Estimating the difficulty of music pieces aids in organizing large collections for music education purposes. However, manually assigning difficulty levels is laborious and might lead to subjective errors [1]. To address this, Music Information Retrieval (MIR) research has focused on automating this process for piano works represented in various modalities [2–6] as well as repertoires from other instruments [7, 8]. Furthermore, the interest of companies like Muse Group [9, 10] and Yousician [11] highlights the industry's recognition of the importance of the task.

Previous work in this field has mainly focused on processing machine-readable symbolic scores [1–4, 12–15]. These, unlike acoustic features extracted from audio whose understanding depends on signal processing knowledge, are both analyzable by computers and interpretable by humans. Musicians find also easier to understand symbolic features since based on music theory knowledge. Initial works towards interpretable difficulty assessment focused on visualization [12], with Chiu and Chen [13] making the first attempt to classify difficulty in the piano repertoire with explainable descriptors. Still, the continually increasing trend towards deep-learning based solutions [3,4], whose lack of transparency limits users' understanding and therefore leads to an eventual non-acceptance in real life applications [16], can impair a fruitful implementation of such technologies in music educational practices.

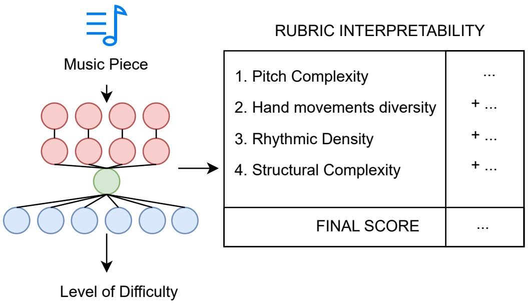


Figure 1: To promote a more objective and transparent assessment, in our white-box model *RubricNet*, similarly as educational rubrics, scores (here difficulty) are dependent on descriptors' values. The Rubric Interpretability table displayed at the right is inspired by [17, Fig. 1]

With this background, we propose a white-box [18] model (cf. Figure 1), which through the concept of a rubric, i. e., an evaluation instrument from music education used to support objective assessment [19–22], allows a transparent interpretation of music difficulty. From this point forward, the white-box model will be denoted by *RubricNet*. Furthermore, to gain a profound understanding of what music difficulty means from an explainable perspective, we build upon the descriptors of Chiu and Chen [13] by proposing a new one focusing on music repetitive patterns. We also provide an interactive companion page [1] to visualize the evaluated data and scrutinize the results in light of its interpretability from a musical point of view.

Through eXplainable Artificial Intelligence (XAI), we

[1] At: https://pramoneda.github.io/rubricnet

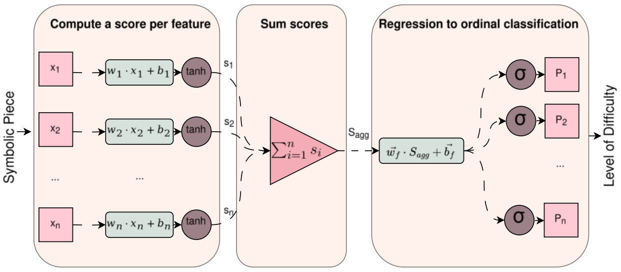


Figure 2: Detailed *RubricNet*'s architecture.

aim to contribute to music education by facilitating the understanding of measurable factors that determine a piece's difficulty. Our work builds on methods from music education, where objectively assessing abstract competences through measuring concrete criteria is consolidated by employing rubrics, i. e., tools which, unlike a black-box, break down complex concepts into simpler ones [19, 22].

Our interpretable methodology aims to bridge the gap between computational models and practical music education needs, enabling educators to make facilitated, but also informed decisions about curriculum development based on the difficulty levels of pieces. We release all the code and models of this research [2], in order to offer a baseline for further research in music difficulty assessment.

2. RELATED WORKS

Previous research aiming to automatically assess the difficulty of piano repertoire examined the link between fingering patterns and the pieces' difficulty level [2, 14, 15]. Recent studies [1, 3, 4, 23] have also made significant contributions. In [3], representations are used to feed three deep learning models—covering music notation, physical gestures, and expressiveness—to emulate Cook's dimensions [24]. These models' predictions are merged using an ensemble method to estimate the scores' difficulty. While we appreciate their musicology-inspired approach, its lack of interpretability harms its usability.

Difficulty estimation of piano pieces has also been investigated through hybrid methods that merge features with deep learning models [1, 23]. However, the absence of publicly shared data and code complicates performing comparative analyses with reference to these works. In [23], the authors combine the methods from Chiu and Chen [13] with deep learning models trained using piano roll as input. In a similar vein, [1] uses JSymbolic features [25] and deep learning models on a proprietary dataset.

In the study by Chiu and Chen [13], 159 pieces from the *8notes* website were used, whereas [23] utilized 1800 MIDI files from the same source. The categorization of these pieces, provided by users of *8notes*, raises concerns about their reliability. Unfortunately, neither study provides access to their data or details on how they were segmented. Other work [26] has attempted to understand the effectiveness of various features, including those proposed in [13] for categorizing the grade levels of a specific piano curriculum. Recent efforts by Zhang et al. [4] and Ramoneda et al. [3] have focused on compiling datasets with difficulty annotations from the established piano publisher Henle Verlag, with the latter's dataset not only being the most extensive but also the only one made publicly available. Therefore, for our comparative analysis, we will use the open-source datasets presented in [3], namely *Can I Play It?* (CIPI), which has 9 levels of difficulty, and *Mikrokosmos-difficulty* (MKD), which includes 3 levels.

Finally, in order to validate our approach, we consider a different and established feature set, i.e., the standard music symbolic features available through Music21 library [27], which includes (amongst others) established JSymbolic features [25], thus facilitating a meaningful comparison with our proposed descriptors. In addition, we also contrast the results achieved with our novel descriptors with those obtained with the features by Chiu and Chen [13], which are also reimplemented and open-sourced in this study. Note that none of the approaches previously mentioned has focused on the interpretability of the descriptors, which is a key contribution of our work.

3. INTERPRETABLE *RubricNet*

The *RubricNet* model (cf. Figure 2) is designed to provide interpretability akin to a rubric, enabling its analysis and results to be intuitively aligned with established practices in music education. This approach ensures that the model's logic and outcomes are easily comprehensible, facilitating their usage in music education along to traditional tools.

3.1 Model Architecture

The network, comprising a series of linear layers dedicated to process individual input descriptors and followed by a nonlinear activation function, is formulated as follows:

Given a set of N input descriptors, each descriptor x_i is first processed through its dedicated linear layer with weight w_i and bias b_i, followed by a hyperbolic tangent activation function to yield:

$$s_i = \tanh(w_i \cdot x_i + b_i) \tag{1}$$

where s_i represents the processed score for the i-th descriptor. Scores are then aggregated in a single score S_{agg}:

$$S_{agg} = \sum_{i=1}^{n} s_i \tag{2}$$

The aggregated score S_{agg} is then passed through a final linear layer to obtain the logits for the class predictions, which are mapped to probabilities with a sigmoid function:

$$\vec{P} = \sigma(S_{agg} \cdot \overrightarrow{w_f} + \overrightarrow{b_f}) \tag{3}$$

where σ denotes the sigmoid function, $\overrightarrow{w_f}$ and $\overrightarrow{b_f}$ are the weight and bias of the final linear layer, respectively.

3.2 Ordinal Optimization

This model applies an ordinal optimization approach [28], predicting ordered categorical outcomes, i. e., difficulty

[2] At: https://github.com/pramoneda/rubricnet

Descriptor	Explanation
Pitch Entropy	Indicates pitch variety; higher values mean more diverse pitch collection
Pitch Range	Distance between the lowest and highest notes.
Average Pitch	Indicating the central pitch level.
Displacement Rate	Measures hand movement intensity across keys reflecting physicality in performance.
Average IOI	The average timing between note onsets, indicative of rhythmic density.
Pitch Set LZ	Indicative of structural complexity and repetitiveness within a pitch set sequence.

Table 1: Explanation of descriptors in musical terms.

levels such as beginner (1), intermediate (2), and advanced (3), through logits. These logits, computed using a mean squared error (MSE) loss, indicate the model's predictions on the ordinal scale. Difficulty level is then obtained as:

$$\max\{i \text{ where } P_i \geq 0.5 \text{ and } P_j \geq 0.5, \forall j < i\} \quad (4)$$

3.3 Interpretability

In *RubricNet*, the *descriptors* (automatically computed from the data) are, to some extent, comparable to the formalized *evaluation criteria* defined in traditional rubrics; similarly, the *aggregated score*, might be comparable to a final *grade/mark* assigned in an educational scenario. Given the correspondences between both, we could consider the model a "white-box" approach, able to promote transparency and interpretability, similarly to a rubric.

It uses independent linear transformations on input descriptors to generate scores between -1 and 1, which directly influence the regression output, S_{agg}. Since negative scores, might be not fully understood in terms of difficulty level, we normalize scores between 0 and 1, rescaling S_{agg} between 0 and 12. This approach mirrors rubric's ability to provide objective and structured feedback, with the simplicity of these transformations aiding in understanding the impact of features in predictions.

The interpretability of the model lies in its ability to dissect each descriptors' influence on a piece's difficulty level. Consequently, analyzing each descriptor's scores might reveal its overall importance on the prediction. Lastly, S_{agg} is a continuous-ordered scalar with rank correlation to difficulty. Therefore, from S_{agg}, we retrieve ordered and discrete categories with clear decision boundaries.

4. EXPLAINABLE DESCRIPTORS

From codified musical scores, we extracted numeric features which are feed to a classification algorithm. We reimplemented a set of features from the literature [13] while proposing a novel one, Pitch Set LZ. In addition to explaining the features (cf. Table 1), we will provide their technical descriptions and analyze their relevance to difficulty and interdependencies using the data.

4.1 Descriptors

In our work, we analyze music sheets encoded in symbolic format, focusing on extracting pitch and timing. Following the approach suggested by Chiu and Chen [13], we process left and right hand parts separately to clarify pedagogical aspects of musical difficulty. Our primary analysis involves sequences of pitch set events, each characterized by a pitch set S and onset time T. Pitch sets, represented by sets of MIDI numbers, are defined over the alphabet of all pitch sets $\mathbf{S}$ that occurred in a score part, while onset times are calculated in seconds from the performance start by the music21 library [27] with reference to marked tempo information. This method emphasizes the timing of note attacks, duration and rests. Additionally, we consider a collection of pitch events, each defined by pitch P over the alphabet of all pitches $\mathbf{P}$. Our analysis started with the five features identified by Chiu and Chen [13] as most relevant to understanding musical difficulty.

Pitch Entropy. The entropy of pitches in the pitch events:

$$-\sum_{i\in\mathbf{P}} p(P=i)\log_2 p(P=i) \quad (5)$$

Pitch Range. The distance between the minimum and maximum MIDI pitches in a score part.

Average Pitch. The average MIDI pitch in a music sheet.

Displacement Rate. Initially proposed by [13], it quantifies the extent of hand movement across the keyboard during the performance of a score. It analyzes maximum pitch distances between consecutive pitch set events and is calculated as a weighted average of three categories: distances less than 7 semitones (assigned a weight of zero); distances over 7 semitones but under an octave (assigned a weight of one); and distances of an octave or larger (assigned a weight of two to emphasize larger movements).

Average IOI: Average Inter Onset Interval. A concept similar to the "Playing speed" introduced by [13], a term we consider deceptive since it actually decreases as the hand's "speed" increases. This is an average time in seconds between onsets of two consecutive pitch set events. Let's denote i^{th} onset time with T_i, then the value is:

$$\frac{\sum_{1\leq i\leq N^{events}-1}(T_{i+1}-T_i)}{N^{events}-1} \quad (6)$$

In 23% of the scores, information about the recommended performance tempo is missing. We then assume the tempo is 100 beats per minute (bpm). Thus, in cases of missing bpm, the Average IOI feature might not be relevant.

Pitch Set LZ. Lempel-Ziv complexity of pitch set sequence. Before introducing our proposed descriptor, it is crucial to provide context and motivation. Pitch Entropy, as emphasized by Chiu and Chen [13], is particularly relevant—a conclusion supported by the analysis of correlations between difficulty and features in the following section, as well as by informal experiments. As Sayood discusses [29], there's a link between entropy of a task and the cognitive load it imposes on the performer, a concept that may also apply to music performance [30]. However, music is often perceived in terms of larger structures like phrases and sections, not just isolated pitches, prompting us to seek a descriptor that captures the "repetitiveness"

Feature	τ_c
Pitch Entropy (R)	0.583
Pitch Set LZ (L)	0.583
Pitch Entropy (L)	0.582
Pitch Set LZ (R)	0.573
Pitch Range (L)	0.567
Pitch Range (R)	0.554
Displacement Rate (R)	0.332
Displacement Rate (L)	0.273
Average IOI (R)	-0.209
Average IOI (L)	-0.208
Average Pitch (R)	0.088
Average Pitch (L)	0.017

Table 2: Features ordered by absolute values of their τ_c rank correlation with the difficulty level.

of music on a broader scale. To this end, we employ LZ-complexity, a measure of redundancy introduced by Lempel and Ziv [31]. In context of music research, it was used for binary encoded rhythm analysis by Shmulevich and Povel [32]. We apply LZ-complexity to sequence of pitch sets: scan a score part, identify all subsequences of pitch sets that cannot be reproduced from preceding material through a recursive copying procedure. The number of such unique subsequences is defined as the LZ-complexity of the part. This approach allows us to assess the structural complexity and redundancy of a musical piece, highlighting the cognitive demands placed on performers.

4.2 Feature Analysis

We assume that, for easier interpretability, features must on average change monotonically with the difficulty level. To measure this quality, we use the τ_c version of Kendall rank correlation coefficient due to its ability to deal with "heavily tied" rankings [33] (many musical pieces have the same difficulty, hence, we have multiple ties in the ranking by difficulty). τ_c is equal to 1 when feature and difficulty rankings are perfectly aligned in the same direction, -1 if they are aligned in opposite directions. As the number of nonconcordant cases increases, the coefficient approaches zero. In Table 2, the results show that the features related to pitch organization are the most correlated to difficulty. Hand displacement and Inter-onset intervals are less correlated, while average pitch seems almost irrelevant.

In addition, we aim to uncover dependencies among the features themselves while mitigating the influence of difficulty, with whom most features are correlated. To achieve this, we calculate conditional τ_c correlations for all feature pairs given a fixed difficulty level, and average the coefficients across all difficulty levels. We then convert these coefficients into a distance matrix and apply hierarchical agglomerative clustering based on average distance to identify clusters of correlated features. From the resulting dendrogram (cf. Figure 3), we observe that features correlated with difficulty—namely Pitch Entropy, Pitch Set LZ, and Pitch Range—are also interrelated. This is remarkable because the three most correlated features are not inherently dependent: one could envision a music piece with any of them maximized while maintaining low values for the others. However, pieces in CIPI typically exhibit coordinated values in these descriptors. Thus, we mostly observe the combined effect of these features, making it challenging to reliably decompose "difficulty" into an aggregate of independent components.

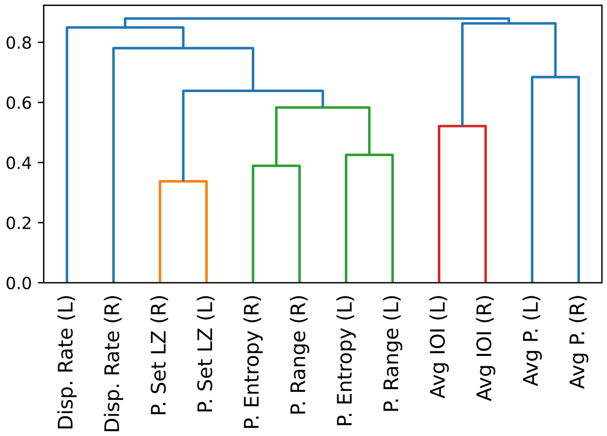


Figure 3: Hierarchical clustering of features based on their average correlation distance within each difficulty class.

5. EXPERIMENTS

5.1 Experimental Setup

To evaluate the effectiveness of our proposed method, we utilized the *Mikrokosmos-difficulty* (MKD), and *Can I play it?* (CIPI) datasets [3]. For fair comparison, we use the 5-fold cross-validation approach defined in [3]. In each split, 60% of the data is used as a train set, while the remaining is equally divided into validation and test sets.

As in [3], we employ mean squared error (MSE) and accuracy within n classes (Acc-n) for evaluation. These metrics are chosen for their applicability to ordinal classification challenges, with Acc-n assessing the model's accuracy for n classes from the true labels, and MSE measuring the average squared prediction error across classes. The effects of dataset imbalances and a fair evaluation across classes are mitigated by macro-averaged metrics.

We optimize the models during training through Adam optimizer with a learning rate of 10^{-2}. The training process incorporates early stopping, based on the Acc-n and MSE metrics from the validation set, to prevent overfitting. Through Ordinal Loss, we frame difficulty prediction as an ordinal classification task, as mentioned in Section 3. We apply a standard scaler and dropout to the features to prevent individual ones from dominating. For each experiment, we look for the best hyperparameters using Bayesian optimization [34]: batch size within the range from 16 to 128, dropout rate between 0.1 and 0.5, learning rate decay from 0.1 to 0.9, and the learning rate itself, tested over a logarithmic scale from $1e-5$ to $1e-1$. This approach allows us to systematically explore the hyperparameter space and identify the optimal settings for our models; thus, enabling a fair comparison between experiments.

5.2 Experimental Results

In Table 3, the results from the comparison between the performance of our novel approach with the presented de-

	CIPI		MKD
	Acc-9	MSE	Acc-3
argnn [3]	32.6(2.8)	2.1(0.2)	75.3(6.1)
virtuoso [3]	35.2(7.3)	2.1(0.2)	65.7(7.8)
pitch [3]	32.2(5.9)	1.9(0.2)	74.2(9.2)
ensemble [3]	39.5(3.4)	**1.1(0.2)**	76.4(2.3)
Ours	**41.4(3.1)**	1.7(0.5)	**79.6(8.8)**

Table 3: Experiment comparison of previous individual deep learning models [3], their ensemble and our explainable and interpretable method on CIPI and MKD.

Experiment	Acc-9	MSE
RubricNet proposed	**41.4(3.1)**	1.7(0.5)
"" with Chiu and Chen [13] descriptors	36.2(5.2)	1.7(0.3)
"" with Music21 descriptors	36.7(6.0)	1.3(0.2)
"" with ALL descriptors	38.9(4.3)	**1.3(0.1)**
"" proposed without Avg P.	39.0(5.6)	1.5(0.4)
"" with positive scores	38.5(3.5)	1.6(0.6)
"" without ordinal regression	36.2(1.3)	2.1(0.4)
Logistic regression	40.0(4.3)	1.5(0.3)

Table 4: Ablation study results for different feature sets (5 first rows) and model configurations (last 3 rows) on CIPI.

scriptors (cf. Sections 3 and 4) and the results achieved by three previous models from the literature (argnn [35], virtuoso [36], pitch) as well as their collective ensemble, are shown. Our model achieves the highest Acc-9 score of $41.4(\pm 3.1)$ in CIPI, surpassing the ensemble's $39.5(\pm 3.4)$, while displaying the second lower MSE of $1.7(\pm 0.5)$, only overtaken by the ensemble's $1.1(\pm 0.2)$. With an Acc-3 score of $79.6(\pm 8.8)$ in the MKD dataset, our approach is superior to previous ones but with a higher standard deviation.

In the following, we examine the impact of various feature and model configurations on *RubricNet* performance (cf. Table 4). As baseline for comparison, we consider the configuration previously discussed (cf. Ours in Table 3).

Employing only the five Chiu and Chen [13] descriptors, i. e., excluding Pitch Set LZ, leads to a decrease in Acc-9 by -5.2, reflecting a performance drop from the baseline. The use of Music21 [27] descriptors, which include JSymbolic [25] and other descriptors widely used in the community, results in a decrease in Acc-9 by -4.7 and a decrease in MSE by -0.4, showing slight improvements in MSE but not in accuracy. However, note that a larger number of descriptors could decrease the explainability. Combining all the descriptors slightly decreases Acc-9 and MSE: -2.5 and -0.4, respectively; with the accuracy results still under the baseline. These results indicate that the descriptors discussed in Section 4 constitute the best option for difficulty estimation on CIPI. Since average pitch showed no relation to difficulty in the feature analysis, we repeated the experiments without this feature. This lead, however, to non-significant worsening of the results.

Concerning the impact of different model configurations, we replace the tanh by sigmoid non-linearities to guarantee positive scores. The obtained MSE rate is similar but the accuracy drops by -3.1. This means that negative scores could aid in training, which is why we keep them, but normalize the scores after the training phase. Besides, substituting the ordinal encoding used in the baseline with a traditional one-hot encoding with cross-entropy loss, results in a decrease in Acc-9 by -5.2 and an increase in MSE by $+0.4$, highlighting the importance of ordinal regression in achieving lower MSE rates. Lastly, logistic regression with ordinal loss decreases the Acc-9 by -1.4 while showing a decrease of MSE by -0.2. This offers a compromise for both metrics but without beating our setup and to our understanding, being less interpretable.

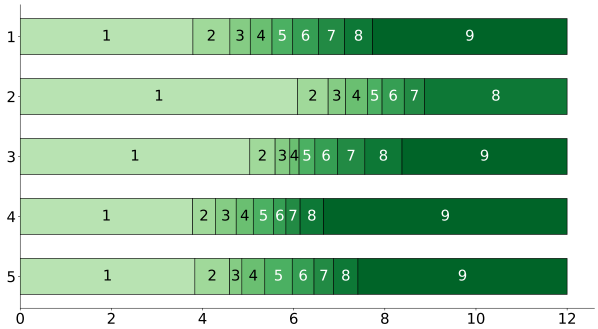


Figure 4: Decision boundaries of the model between grades on S_{agg} (X axis) for all splits (Y axis) on CIPI.

Overall, the gains offered by Rubricnet with the features proposed are relatively modest compared to the baselines. However, having a smaller feature set is necessary for explainability. The novelty of our approach lies in aligning the interpretability of music education with rubric-like interpretability feedback. This alignment is essential for a successful application of our model in practice, as we will discuss in further sections.

5.3 Decision Boundaries

In *RubricNet*, the input features are combined into a single scalar before performing the final ordinal classification. Analysis of the results shows that the final layer defines optimized decision boundaries, setting thresholds for S_{agg} that progressively increase along with difficulty levels. Because of the final sigmoid activation, once S_{agg} exceeds a boundary, the corresponding difficulty level will always be active, which guarantees the ordinality of the predictions. By examining the decision boundaries (cf. Figure 4), we observe that the trends are similar across splits, displaying shorter valid ranges around intermediate levels. Note that in split 2, there are only 8 classes because the model ignored the last class. This can happen as we use numeric optimization, which sometimes falls into local minima. These minima might seem optimal based on the validation metrics but do not meet our overall performance expectations.

6. DISCUSSION AND LIMITATIONS

Now, we analyze whether *RubricNet* is interpretable from a musical point of view. To understand how features impact the final level suggested by the model, we evaluate the contribution of each descriptor to the aggregated score. Since learning to play an instrument is a progressive process, relative contributions of features to different levels

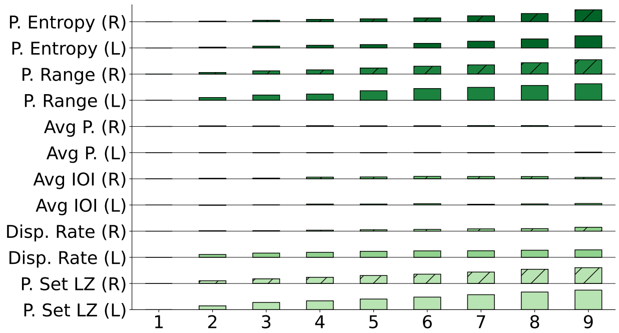
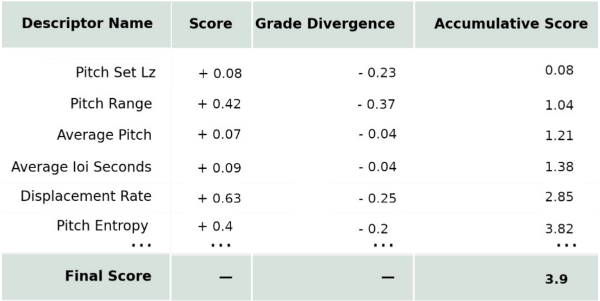


Figure 5: Average relative contribution of descriptors (Y axis) normalized between 0 and 1, across grades (X axis).



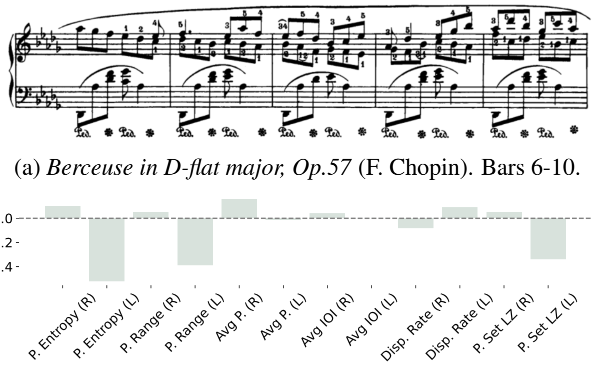
Descriptor Name	Score	Grade Divergence	Accumulative Score
Pitch Set Lz	+ 0.08	- 0.23	0.08
Pitch Range	+ 0.42	- 0.37	1.04
Average Pitch	+ 0.07	- 0.04	1.21
Average Ioi Seconds	+ 0.09	- 0.04	1.38
Displacement Rate	+ 0.63	- 0.25	2.85
Pitch Entropy	+ 0.4	- 0.2	3.82
...	...	...	...
Final Score	—	—	**3.9**

Figure 6: Simplified difficulty interpretable rubric for the *Nocturne op. 9, no. 3* (F. Chopin). Descriptors' values for the right hand and final score (for both hands) are shown.

with reference to the grade 1 are displayed instead of absolute values. These contributions are averaged across splits on the test set and shown in Figure 5.

We observe a trend of higher contributions when the level increases for every descriptor. This observation is consistent with the fact that S_{agg} value increases for higher levels (cf. Figure 4). The most discriminative features are pitch entropy and pitch range, as well as the LZ descriptor for higher levels. Conversely, some features, e. g., average IOI or the average pitch, have low contribution to the model's decisions, as shown by their relatively constant and small values across grades. The latter is expected, since very different pieces could have the same average pitch, not disclosing anything about difficulty. The former might be explained by the averaging, which can remove information, especially when a piece can alternate between fast and slow parts. Besides, as mentioned before, tempo is often poorly annotated in the dataset.

To better understand the explainable capabilities of the proposed descriptors, in the following, we provide a musical examination of two concrete samples, by this demonstrating the interpretability of our approach. *Nocturne op. 9, no. 3* by F. Chopin, is labelled as level 7, but classified as level 2. All the descriptors are below the grade average, as shown in the rubric (cf. *Grade Divergence* in Figure 6). Our hypothesis is that this nocturne contains many challenges that go beyond the descriptors used. There are constant changes in dynamics, a variety of articulations, and as a key difficulty aspect, many types of polyrhythms between the right and left hands. Further research should address all the types of difficulty challenges, probably underrepresented in the existing datasets.



(a) *Berceuse in D-flat major, Op.57* (F. Chopin). Bars 6-10.

(b) Distance to the average for the grade of the scores. Extracted from original rubric (grade divergence column).

Figure 7: Musical excerpt (a) and a rubric outcome (grade divergence) plotted (b) from a piece in level 7.

The piece *Berceuse in D-flat major, Op.57* by F. Chopin, shown in Figure 7 is appropriately classified as grade 7. This is because it maintains a left-hand accompaniment with few changes, in contrast to the higher virtuosity of the right hand. The left hand has scores below average in most descriptors because of its few changes: Pitch Range (-0.41), Average IOI (-0.04), Pitch Set LZ (-0.34), Average Pitch (-0.01), and Pitch Entropy (-0.52). In contrast, the right hand shows more virtuosity, with higher than average scores for all the right hand features. These scores collectively contribute to a final cumulative score that accurately reflects the overall difficulty.

Finally, it should be noted that our approach primarily focuses on descriptors related to pitch sequences and onsets, while disregarding others. Still, the ablation study showed that other features sets (e. g., those from music21), even covering aspects like rhythm variety, do not enhance our classifier's performance either. In addition, expressive elements [37] such as dynamics, tempo changes, and articulation, since often left to performers' interpretation, are not always captured in musical notation [3], and therefore is a dimension our score-based model does not consider.

7. CONCLUSION AND FUTURE WORK

In our study, we proposed a novel white-box parameter-efficient model aligned with the music education community tools, i. e., rubrics, which outperforms previous approaches on difficulty estimation. In addition, we created an interactive companion page for visualizing CIPI and MKD datasets. In summary, we showed that analyzing explainable descriptors, unlike deep learning models, offers clarity, which gives both teachers and students specific insights into pieces. This approach not only underscores the importance of explainable artificial intelligence (XAI) in understanding music difficulty, but also emphasizes the potential for such technologies to contribute to the broader field of music education. For future research, we consider interesting to creating a dataset based on technical challenges like finger fluency and polyphonic complexity, as well as user studies for understanding the perception of interpretable feedback by music education community.

8. ETHICS STATEMENT

The system presented in this paper aims at obtaining the difficulty of a musical piece through several descriptors. In previous work, descriptors were not available, limiting access to the area. This situation underscores the need for open science practices. Therefore, we open our implementation, to facilitate access for new researchers. Besides, the dataset used for this study is available upon request for non-profit and academic research purposes. While this limits its use in commercial applications, it ensures the reproducibility of the results. The data consists of open-source scores of music that is no longer copyrighted, its use for open research can thus be considered fair.

The proposed work belongs to the area of assisted music learning. One might argue that such a tool can have a detrimental impact on music teaching jobs. While this is a valid concern, we think that an eventual solution of the addressed task, would not endanger music educators profession, whose role naturally goes much beyond than categorizing music in difficulty levels. Instead, this technology should be seen as a way to support them in the own teaching practices, for instance, by alleviating their burden on some duties, such as exploring large collections, and by this enabling them to easily discover forgotten musical works from our cultural heritage which fit students' needs. Moreover, through this research, we also aim to convey the message that the path to advancement does not solely lie in acquiring more data or creating larger models. By highlighting what drives its decisions, our proposed model aligns with the goals of eXplainable AI, something crucial for its acceptance in music education. Although our efforts in making the system interpretable and explainable will partly answer the common criticisms made to black-box approaches, the real impact of our system remains to be verified by its future use in real scenarios.

9. ACKNOWLEDGEMENTS

We want to thank Alia Morsi's previous work on difficulty estimation from a feature engineering perspective, which encouraged research in tabular data [26]. The team would also acknowledge Marius Miron for continuously insisting on aligning difficulty estimation with explainability. He also highlighted the direction of using a rubric-like explainability feedback, pointing out Ustun et al.'s work [17]. We also thank Roser Batlle-Roca for helping to discuss some concepts between interpretability and explainability based on her research [38].

This work was supported by "IA y Música: Cátedra en Inteligencia Artificial y Música" (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, and the European Union-Next Generation EU, under the program "Cátedras ENIA 2022 para la creación de cátedras universidad-empresa en IA". This work was also supported by the ANR project TABASCO (ANR-22-CE38-0001) and the travel grant MERMOZ2-012047. Finally, this work was also possible through the support of the Hightech Agenda Bayern, funded by the Free State of Bavaria (Germany).

10. REFERENCES

[1] D. S. Deconto, E. L. F. Valenga, and C. N. Silla, "Automatic music score difficulty classification," in *Proc. of the 30th IEEE Int. Conf. on Systems, Signals and Image Processing (IWSSIP)*, Ohrid, North Macedonia, 2023.

[2] P. Ramoneda, N. C. Tamer, V. Eremenko, M. Miron, and X. Serra, "Score difficulty analysis for piano performance education," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022.

[3] P. Ramoneda, D. Jeong, V. Eremenko, N. C. Tamer, M. Miron, and X. Serra, "Combining piano performance dimensions for score difficulty classification," *Expert Systems with Applications*, vol. 238, pp. 1–16, 2024.

[4] H. Zhang, E. Karystinaios, S. Dixon, G. Widmer, and C. E. Cancino-Chacón, "Symbolic music representations for classification tasks: A systematic evaluation," in *Proc. of the 24th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Milan, Italy, 2023.

[5] P. Ramoneda, D. Jeong, J. J. Valero-Mas, and X. Serra, "Predicting performance difficulty from piano sheet music images," in *Proc. of the 19th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Milano, Italy, 2023.

[6] P. Ramoneda, M. Lee, D. Jeong, J. J. Valero-Mas, and X. Serra, "Can audio reveal music performance difficulty? insights from the piano syllabus dataset," *arXiv preprint arXiv:2403.03947*, 2024.

[7] M. A. V. Vásquez, M. Baelemans, J. Driedger, W. Zuidema, and J. A. Burgoyne, "Quantifying the ease of playing song chords on the guitar," in *Proc. of the 24th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Milan, Italy, 2023.

[8] E. Holder, E. Tilevich, and A. Gillick, "Musiplectics: Computational assessment of the complexity of music scores," in *Proc. of the ACM Int. Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward!)*, Pittsburgh, USA, 2015.

[9] "Musescore have automatic difficulty categories from year 2022," https://musescore.com/, accessed on April 12, 2024.

[10] "Ultimate guitar have automatic difficulty categories from year 2022," https://www.ultimate-guitar.com/, accessed on April 12, 2024.

[11] "System for estimating user's skill in playing a music instrument and determining virtual exercises thereof," Patent US9 767 705B1, 2017.

[12] V. Sébastien, H. Ralambondrainy, O. Sébastien, and N. Conruyt, "Score analyzer: Automatically determining scores difficulty level for instrumental e-learning,"

in *Proc. of the 13th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Porto, Portugal, 2012.

[13] S.-C. Chiu and M.-S. Chen, "A study on difficulty level recognition of piano sheet music," in *Proc. of the IEEE Int. Symposium on Multimedia (ISM)*, Irvin, USA, 2012.

[14] E. Nakamura, N. Ono, and S. Sagayama, "Merged-output hmm for piano fingering of both hands." in *Proc. of the 15th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Taipei, Taiwan, 2014.

[15] E. Nakamura and S. Sagayama, "Automatic piano reduction from ensemble scores based on merged-output hidden markov model," in *Proc. of the 41st Int. Computer Music Conf. (ICMC)*, Denton, USA, 2015.

[16] D. Branley-Bell, R. Whitworth, and L. Coventry, "User trust and understanding of explainable ai: Exploring algorithm visualisations and user biases," in *Proc. of the Int. Conf. on Human-Computer Interaction (HCII)*, Copenhagen, Denmark, 2020.

[17] B. Ustun and C. Rudin, "Learning Optimized Risk Scores," *Journal of Machine Learning Research*, vol. 20, no. 150, pp. 1–75, 2019.

[18] O. Loyola-Gonzalez, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE access*, vol. 7, pp. 154 096–154 113, 2019.

[19] M. E. Latimer, M. J. Bergee, and M. L. Cohen, "Reliability and perceived pedagogical utility of a weighted music performance assessment rubric," *Journal of Research in Music Education*, vol. 58, pp. 168 – 183, 2010.

[20] M. Álvarez-Díaz, L. M. Muñiz-Bascón, A. Soria-Alemany, A. Veintimilla-Bonet, and R. Fernández-Alonso, "On the design and validation of a rubric for the evaluation of performance in a musical contest," *International Journal of Music Education*, vol. 39, pp. 66 – 79, 2020.

[21] B. C. Wesolowski, "Understanding and developing rubrics for music performance assessment," *Music Educators Journal*, vol. 98, pp. 36 – 42, 2012.

[22] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity, and educational consequences," *Educational Research Review*, vol. 2, pp. 130–144, 2007.

[23] Y. Ghatas, M. Fayek, and M. Hadhoud, "A hybrid deep learning approach for musical difficulty estimation of piano symbolic music," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 1–14, 2022.

[24] N. Cook, "Analysing performance and performing analysis," *Rethinking Music*, vol. 8, pp. 1–23, 1999.

[25] I. F. McKay, Julie E. Cumming, "jSymbolic 2.2: Extracting features from symbolic music for use in musicological and MIR research." in *Proc. of the 19th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Paris, France, 2018.

[26] A. Morsi, "Characterizing difficulty levels of keyboard music scores," Master's thesis, Music Technology Group, Universitat Pompeu Fabra, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4090526

[27] M. S. Cuthbert and C. Ariza, "Music21: A toolkit for computer-aided musicology and symbolic music data," in *Proc. of the 11th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Utrecht, Netherlands, 2010.

[28] J. Cheng, Z. Wang, and G. Pollastri, "A neural network approach to ordinal regression," in *Proc. of the IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, Hong Kong, China, 2008.

[29] K. Sayood, "Information theory and cognition: A review," *Entropy*, vol. 20, pp. 1–19, 2018.

[30] C. Palmer, "The nature of memory for music performance skills," in *Music, Motor Control and the Brain*, E. Altenmüller, J. Kesselring, and M. Wiesendanger, Eds. Oxford, UK: Oxford University Press, 2012.

[31] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[32] I. Shmulevich and D.-J. Povel, "Measures of temporal pattern complexity," *Journal of New Music Research*, vol. 29, no. 1, pp. 61–69, 2000.

[33] M. Kendall, *Rank Correlation Methods*, ser. Griffin books on statistics. Griffin, 1962.

[34] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data mining (KDD)*, Anchorage, USA, 2019.

[35] P. Ramoneda, D. Jeong, E. Nakamura, X. Serra, and M. Miron, "Automatic piano fingering from partially annotated scores using autoregressive neural networks," in *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, Lisboa, Portugal, 2022.

[36] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam, "VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[37] 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)*. IEEE, 2023, pp. 1–5.

[38] R. Batlle-Roca, E. Gómez, W. Liao, X. Serra, and Y. Mitsufuji, "Transparency in music-generative ai: A systematic literature review," 2023. [Online]. Available: http://dx.doi.org/10.21203/rs.3.rs-3708077/v1