PREDICTING PERFORMANCE DIFFICULTY FROM PIANO SHEET MUSIC IMAGES

Pedro Ramoneda[1] **Jose J. Valero-Mas**[1]
Dasaem Jeong[2] **Xavier Serra**[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona
{pedro.ramoneda, josejavier.valero, xavier.serra}@upf.edu
[2] MALer Lab, Sogang University, Seoul
dasaemj@sogang.ac.kr

ABSTRACT

Estimating the performance difficulty of a musical score is crucial in music education for adequately designing the learning curriculum of the students. Although the Music Information Retrieval community has recently shown interest in this task, existing approaches mainly use machine-readable scores, leaving the broader case of sheet music images unaddressed. Based on previous works involving sheet music images, we use a mid-level representation, bootleg score, describing notehead positions relative to staff lines coupled with a transformer model. This architecture is adapted to our task by introducing an encoding scheme that reduces the encoded sequence length to one-eighth of the original size. In terms of evaluation, we consider five datasets—more than 7500 scores with up to 9 difficulty levels—, two of them particularly compiled for this work. The results obtained when pretraining the scheme on the IMSLP corpus and fine-tuning it on the considered datasets prove the proposal's validity, achieving the best-performing model with a balanced accuracy of 40.34% and a mean square error of 1.33. Finally, we provide access to our code, data, and models for transparency and reproducibility.

1. INTRODUCTION

Estimating the difficulty of a piece is crucial for music education, as it enables the effective structuring of music collections to attend to the student's needs. This has led to a growing research interest [1–4], as well as the development of automatic systems for exploring difficulties by major industry players such as Muse Group [5,6] and Yousician [7].

Previous research on predicting piano difficulty has primarily focused on symbolic machine-readable scores [1, 2, 4, 8–10]. Early studies explored feature engineering descriptors [1,2] and the relationship between piano fingering and difficulty [8–10]. A recent study [4] used stacked recurrent neural networks and context attention for difficulty classification on machine-readable scores, employing embeddings from automatic piano fingering, piano expressive generation [11], and score information. This study found that modeling the score difficulty classification task as an ordinal regression problem [12] was advantageous, and using entire pieces for training, rather than fragments, was essential to avoid degraded performance.

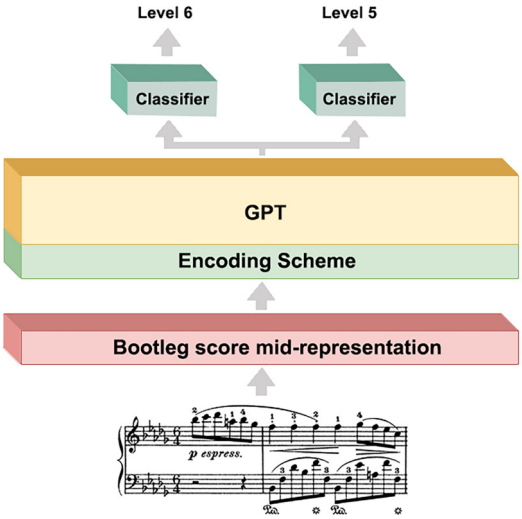

Figure 1. We consider the bootleg score mid-representation with a multi-task GPT-based recognition framework to predict the performance difficulty associated to a piano score directly from sheet images from multiple annotated collections with varied difficulty levels.

Although symbolic machine-readable scores offer more interpretability [10], with all the music information completely accessible, their limited availability compared to sheet music images restricts the practical use of difficulty prediction tools for librarians, teachers, and students. Focusing on sheet music image analysis expands the range of available music, has the potential to preserve the cultural heritage of symbolic-untranscribed scores, and addresses the lack of diversity in Western classical piano curricula. By analyzing image-based sheet music, we aim

to create technology for highlighting historically underrepresented communities like female composers [13, 14] and promoting diversity in piano education. This promotion is crucial since the piano teaching repertoire has remained mostly unchanged for decades [15], containing around 3,300 pieces [16], while projects such as IMSLP house remarkably larger databases.

One of the main challenges in working with sheet music is attaining a symbolic music-based representation for direct analysis. Although Optical Music Recognition (OMR) literature has considerably improved in creating such representations over the past 30 years, it remains an unsolved task [17]. Bootleg score [18] is an alternative to symbolic scores obtained with OMR. This mid-level symbolic representation keeps the most relevant primitives of the music content in a music sheet, which has shown remarkable success in several tasks [19–22], especially in classification, such as piano composer classification [19, 23, 24] or instrument recognition [25].

We build on this literature, employing the GPT model [26] and bootleg score in our analysis. More precisely, we consider the approach by Tsai et al. [18], in which a GPT model pretrained on the IMSLP piano collection is finetuned for specific recognition tasks. With adequate adaptations, we hypothesize that this framework may also succeed in estimating performance difficulty on music sheet images.

As aforementioned, difficulty estimation benefits from the use of entire music pieces rather than excerpts to obtain adequate success rates. However, processing large sequence stands as a remarkable challenge in music processing, especially when addressing bootleg representations due its considerable verbosity. While some recent mechanisms address this issue in general learning frameworks (*e.g.*, Flash Attention [27]), we extend the original proposal by Tsai et al. [18] with a multi-hot optimization target for GPT pretraining, and replace the categorical encoding with causal convolutional or feedforward projection layers to enhance performance and reduce costs.

Moreover, addressing data scarcity is crucial for promoting and establishing this task within the Music Information Retrieval community. As of now, the *Mikrokosmos-difficulty* (MK) [10] and *Can I Play It?* (CIPI) [4] symbolic datasets stand for the only available annotated collections, out of which music sheet images can be obtained by engraving mechanisms. To enhance data availability and encourage further research, we have collected additional datasets from existing collections, namely *Pianostreet-difficulty* (PS), *Freescore-difficulty* (FS), and black female composers collection Hidden Voices (HV). This results in more than 7500 music pieces, spanning up to 9 difficulty levels and each annotated with a difficulty classification system. Although difficulty prediction contains a subjective element, global trends may emerge when examining multiple difficulty classification systems simultaneously. To our knowledge, no previous research has explored this aspect. Consequently, we propose a multitask approach to training simultaneously on CIPI, PS, and FS datasets. Finally, we also analyze the generalization of our proposed methodologies with the MK and HV benchmark datasets.

Considering all above, our precise contributions are: (i) we adopt the previous bootleg-representation literature [23, 24], pretraining a GPT model on IMSLP and finetuning it for our task, adapting the encoding scheme accordingly, as presented in Figure 1; (ii) we evaluate our proposal using a novel sheet music image collection of five datasets with more than 7,500 pieces with difficulty levels ranging up to 9; (iii) we propose a multi-task strategy for combining multiple difficulty classification systems from the datasets; (iv) we conduct extensive experiments to assess the proposed methodologies, including a zero-shot scenario for testing generalization and comparisons with previous proposals on the CIPI dataset; and (v) to promote the task, code, and models [1], and datasets [2] are publicly available.

2. MUSIC SHEET IMAGE DATASETS

Due to the relative recentness of the field, the lack of annotated corpora has severely constrained the performance difficulty assessment. The earliest data assortments may be found in the works by Sebastian et al. [1] and Chiu et al. [2], which respectively collected 50 and 300 MIDI scores from different score repositories. However, these datasets were never publicly released.

To our best knowledge, the *Mikrokosmos difficulty* (MK) set by Ramoneda et al. [10], which comprises 147 piano pieces by Béla Bartók in a symbolic format graded by the actual composer, represents the first publicly available collection for the task at hand. More recently, the authors introduced the *Can I Play It?* (CIPI) dataset [4], a collection of 652 piano works in different symbolic formats annotated after 9 different difficulty levels. Note that, while sheet music scores can be obtained by resorting to engraving mechanisms, the insights obtained may not apply to real-world scenarios.

Dataset	Pieces	Classes	AIR	Noteheads	Composers
MK [10]	147	147	.78	49.2k	1
CIPI [4]	652	9	.33	1.1M	29
PS	2816	9	.24	7.2M	92
FS	4193	5	.37	5.8M	747
HV	17	4	1	21.5k	10

Table 1. Description of existing collections for performance difficulty estimation based on the number of pieces, classes, average imbalance ratio (AIR), noteheads, and composers. The dashed line differentiates the datasets based on symbolic (above) and image (below) sheet music.

To address this limitation, we compiled a set of real sheet music images of piano works together with their performance difficulty annotations from different music education and score-sharing platforms on the Internet. More

[1] https://github.com/PRamoneda/pdf-difficulty
[2] https://zenodo.com/record/8126801

precisely, we arranged three different collections attending to the source: (i) the *Pianostreet-difficulty* (PS) set retrieved from [28] that depicts 2,816 works with 9 difficulty levels annotated by the Pianostreet team; (ii) the *Freescores-difficulty* (FS) assortment from [29] that contains 4,193 pieces with 5 difficulty levels comprising a variety of compositions and annotations by the users of the platform; and (iii) the *Hidden Voices* (HV) collection [30, 31], a set of 17 pieces by black female composers annotated with 4-level difficulty labels by musicologists of the Colorado Boulder Music Department.

Table 1 summarizes the main characteristics of commented publicly-available collections. The *average imbalance ratio* (AIR), measured as the mean of the individual ratios between each difficulty class and the majority label in each collection, is also provided for reference purposes.

3. METHODOLOGY

Based on its success when addressing classification tasks from sheet music images [23, 25], our proposal considers the use of the so-called bootleg score representation coupled with a GPT-based recognition model to estimate the performance difficulty of a piece.

Introduced by [18], bootleg scores stand as a simple—yet effective—representation to encode the content of a sheet music image for certain recognition tasks. Formally, a bootleg score is a binary matrix of length w and $h = 62$ vertical positions—*i.e.*, $\mathcal{X} \in \{0,1\}^{w \times 62}$—that respectively denote the temporal and pitch dimensions. Note that the w value represents the number of note heads detected by the bootleg extraction process. Our work resorts to this representation, being the use of alternative codifications posed as a future line to address.

The GPT recognition framework undergoes an unsupervised pretraining step on the IMSLP piano collection, which was originally used by [18]. Eventually, considering a set of labeled data $\mathcal{T} \subset \mathcal{X} \times \mathcal{C}$ where $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ denotes the possible difficulty levels, the model is fine-tuned to retrieve the recognition function $\hat{f} : \mathcal{X} \rightarrow \mathcal{C}$ that relates a bootleg representation to a particular difficulty level. Based on previous work addressing this task [4], we consider an ordinal classification framework [12] as the difficulty grading scales naturally fit this formulation.

Despite being capable of addressing the task, the framework was noticeably affected by two factors: (i) the excessive length of the input sequences when pretraining the model; and (ii) the inconsistent definition of difficulty levels among corpora. Consequently, we introduce two mechanisms specifically devised to address these limitations.

3.1 Sequence length in pretraining

One of the main drawbacks related to bootleg representations is their verbosity, as it depicts $h = 62$ elements per frame. To address this issue, Tsai et al. [23] proposed subdividing each column into groups of 8 elements and encoding each according to a vocabulary of $|\sigma| = 2^8$ elements. In this regard, the initial bootleg score $x \in \{0,1\}^{w \times 62}$ is mapped to a novel space defined as $\Sigma^{w \times 8}$. This representation is then flattened to undergo a categorical embedding process that maps it to a feature-based space denoted as $\mathbb{R}^{8w \times 768}$, which is eventually used for pretraining the GPT model with 768-dim hidden states. Note that this process reduces the vocabulary size and remarkably increases the sequence length.

To address this issue, we propose substituting this tokenization process with an embedding layer that directly maps the bootleg score into a suitable representation, avoiding the extension of the initial length of the sequence. In this sense, the initial bootleg representation $x \in \{0,1\}^{w \times 62}$ is mapped to a space defined as $\mathbb{R}^{w \times 768}$ that serves as input to the GPT model with a fraction of the length of the encoding used by Tsai et al. [23]. Besides reducing the length of the sequences to process, we hypothesize that such an embedding may benefit the recognition model as a suitable representation is inferred for the task. In this regard, our experiments will compare two types of embedding approaches—more precisely, a fully-connected layer and a convolutional one, respectively denoted as FC and CNN—to quantitatively assess this claim.

Figure 2 graphically describes the approach by Tsai et al. [23] and the presented proposal. In opposition to the reference work, the proposal considers multi-hot encoding instead of discrete categorical index as the output of the GPT recognition framework, by using binary cross-entropy loss instead of negative log-likelihood loss.

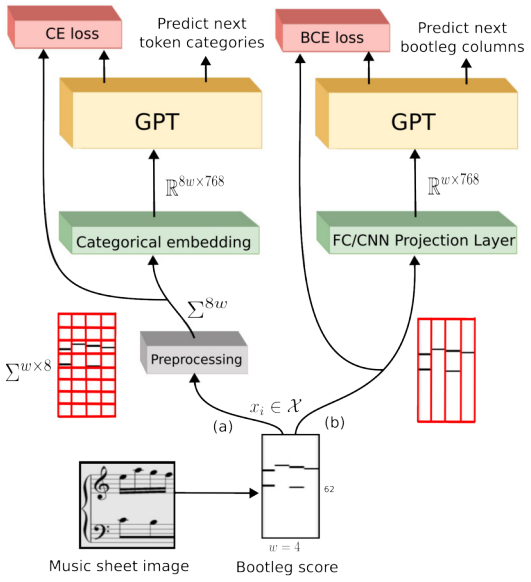

Figure 2. Comparison between the proposal by Tsai et al. [23]—denoted as (a)—and the presented proposal—highlighted as (b)—for a case of toy example with a duration of $w = 4$.

3.2 Multi-task learning of multiple difficulty classification systems

The pretrained GPT model can be simply finetuned for a performance difficulty classification task by adding a projection layer and a learnable classification token, as de-

picted in Figure 3. However, the actual definition of the performance difficulty of a piece is a highly subjective problem that may bias—and, hence, remarkably hinder—the goodness of a recognition model. In this regard, we hypothesize that using a multi-task approach that attends different definitions of difficulty—*i.e.*, a labeled assortment of data from multiple annotators—may benefit the generalization capabilities of the approach.

In this regard, we modify the reference architecture for the downstream task to include an additional classification layer for each training collection. While simple, such a proposal is expected to improve the overall recognition performance given the wider variety of data provided during the training process. Figure 3 graphically describes this proposal.

Finally, no pre-processing is done in relation to the label distribution of the corpora to avoid inducing any type of bias. In this regard, the sampling protocol of the model has been forced to maintain its original distributions.

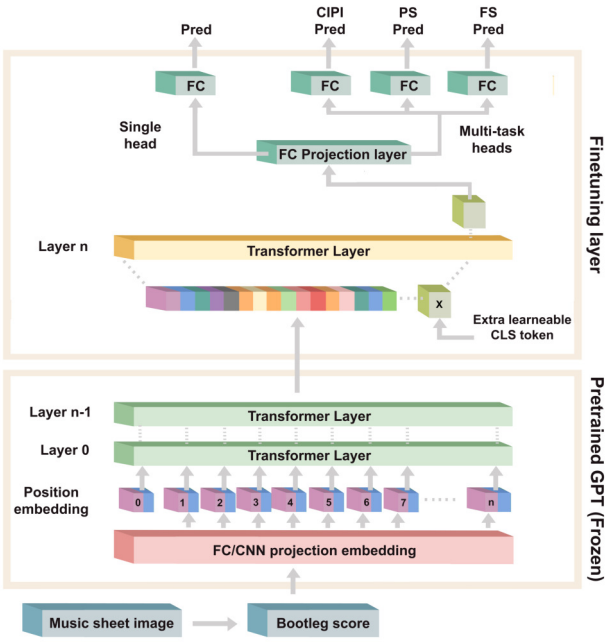

Figure 3. Graphical description of the downstream architecture depicting the classification heads for the multi-task proposals as well as the single-head case of the reference work.

4. EXPERIMENTAL SETUP

4.1 Data collections and assessment metrics

To validate the proposal, we have considered the five publicly-available data collections presented in Section 2, *i.e.*, *Mikrokosmos difficulty* (MK) [10], *Can I Play It?* (CIPI) [4], *Pianostreet-difficulty* (PS) [28], *Freescores-difficulty* (FS) [29], and *Hidden Voices* (HV) [30, 31]. While MK and CIPI exclusively comprise symbolic scores, we engraved them into music sheets and included them due to the commented scarcity of annotated data.

We considered a 5-fold cross-validation scheme with a data partitioning of 60% for the finetuning phase after the pretraining stage with IMSLP together with two equal-size splits of the remaining data for validation and testing. Note that, since MK and HV are exclusively used for benchmark purposes, no partitioning is applied to them.

In terms of performance evaluation, we resort to two assessment criteria typically used in ordinal classification [32]: *accuracy within* n (Acc_n) and *mean squared error* (MSE). To adequately described them, let $\mathcal{S} \subset \mathcal{X} \times C$ denote a set of test data and let $\mathcal{S}_c = \{(x_i, y_i) \in \mathcal{S} : y_i = c\}$ with $1 \leq i \leq |\mathcal{S}|$ be the subset of elements in $\mathcal{S}$ with class c.

Based on this, Acc_n is defined as:

$$\text{Acc}_n = \frac{1}{|\mathcal{C}|} \sum_{\forall c \in \mathcal{C}} \frac{\left| \left\{ y \in \mathcal{S}_c : \left| \hat{f}(x) - c \right| \leq n \right\} \right|}{|\mathcal{S}_c|} \quad (1)$$

where $\hat{f}(\cdot)$ represents the trained recognition model and $n \in \mathbb{N}_0$ denotes the tolerance or class-boundary relaxation that allows for errors in adjacent labels. In our experiments we consider the values of $n = 0$ (no tolerance) and $n = 1$ (smallest adjacency tolerance), respectively denoted as Acc_0 and Acc_1 in the rest of the work.

Regarding MSE, this figure of merit is defined as:

$$\text{MSE} = \frac{1}{|\mathcal{C}|} \sum_{\forall c \in \mathcal{C}} \frac{\sum_{\forall x \in \mathcal{S}_c} \left(\hat{f}(x) - c \right)^2}{|\mathcal{S}_c|} \quad (2)$$

Finally, note that all these metrics are macro-averaged to account for the unbalanced nature of the data collections used in the work.

4.2 Training procedure

As commented, the recognition model undergoes an initial pretraining stage considering the IMSLP corpus. During this stage, the model considers sequences of 256 tokens, each with a binary cross-entropy as a loss function. To speed up this process, the Flash Attention framework by [27] is also considered. For comparative purposes, all other parameters remain unaltered from the reference works [23].

After that, the model is finetuned on the downstream difficulty estimation task, considering an Adam optimizer [33] with a learning rate of 10^{-5} and early stopping based on the Acc_0 and MSE metrics on the validation set. Moreover, a balanced sampler is considered to tackle the issue of unbalanced data collections. Ordinal Loss [12] is applied to train the difficulty prediction as an ordinal classification problem, while no loss weighting considered in the multi-task framework. For regularization and stable training, gradient clipping is set to 10^{-4}, with a batch size of 64 and L2 regularization. This optimization process is carried out exclusively on the last layer of the model, resorting the remaining parts to the weights obtained during the pretraining phase of the procedure.

Note that while these processes may be further studied to account for the optimal solution that retrieves the best-performing results, such a study is out of the scope of the work and is left as future work to address.

5. EXPERIMENTS AND RESULTS

This section presents the results obtained with the introduced experimental scheme. To adequately provide insights about the task, the section provides a series of individual experiments devoted to analyzing one aspect of the proposal: Section 5.1 analyzes the influence of the encoding scheme; Section 5.2 evaluates the influence of the multitask architecture; Section 5.3 delves on the ranking generalization in a zero-shot scenario; finally, Section 5.4 compares the attainable results when addressing the task from the symbolic versus the sheet-image domains.

5.1 Encoding schemes experiment

This first experiment compares the performance of the two encoding schemes presented in Section 3.1, *i.e.*, GPT_{FC} and GPT_{CNN}. Table 2 presents the results obtained for the CIPI, FS, and PS collections for the three figures of merit considered.

Encoding	Acc_0 (%)	Acc_1 (%)	MSE
Can I Play it?			
GPT_{FC}	34.3(6.1)	78.1(4.6)	1.6(0.3)
GPT_{CNN}	**36.2(8.2)**	**81.7(1.5)**	**1.4(0.1)**
PianoStreet			
GPT_{FC}	30.9(3.8)	71.1(9.6)	2.1(0.4)
GPT_{CNN}	**31.8(1.6)**	**78.8(1.8)**	**1.9(0.1)**
FreeScores			
GPT_{FC}	46.6(1.9)	**92.5(1.0)**	**0.8(0.1)**
GPT_{CNN}	**47.3(3.4)**	92.4(0.6)	0.8(0.1)

Table 2. Results of comparing the encoding schemes GPT_{FC} and GPT_{CNN}. Bold values highlight the best results per collection and metric.

As it may be observed, the GPT_{CNN} experiment outperformed the GPT_{FC} experiment in most evaluation metrics across the three datasets. More precisely, the GPT_{CNN} consistently achieved the best performance in the Acc_0 metric for all data collections, showing an average improvement of 1% concerning the GPT_{CNN} case. This trend remains for the rest of the figures of merit except for the case in the FS assortment, in which the results of the FC-based model outperform those of the CNN case.

Nevertheless, attending to the high standard deviations, the performance results of the two models show a remarkable overlap in performance, hence suggesting that both schemes are equally capable of performing the posed task of score difficulty analysis from sheet music images. In this regard, further work should explore other encoding alternatives to assess whether this performance stagnation is due to the representation capabilities of the considered embedding layers or due to the recognition framework.

5.2 Multi-task learning experiment

In this second study, we assess the capabilities of the multi-task framework proposed in Section 3.2 trained simultaneously on the CIPI, PS, and FS datasets for the two $\text{GPT}^{\texttt{multi}}_{\text{FC}}$ and $\text{GPT}^{\texttt{multi}}_{\text{CNN}}$ encoding schemes. Table 3 provides the results obtained.

Encoding	Acc_0 (%)	Acc_1 (%)	MSE
$\text{GPT}^{\texttt{multi}}_{\text{FC}}$			
CIPI	**40.3(4.3)**	**82.0(1.4)**	**1.3(0.1)**
PS	35.9(3.1)	**78.2(3.4)**	**1.9(0.2)**
FS	45.8(2.5)	92.0(1.4)	**0.8(0.1)**
$\text{GPT}^{\texttt{multi}}_{\text{CNN}}$			
CIPI	34.9(5.0)	81.4(1.3)	1.4(0.1)
PS	**35.9(2.8)**	74.5(3.4)	2.7(0.2)
FS	**45.9(1.2)**	**92.4(2.1)**	0.8(0.1)

Table 3. Results of multi-task learning experiment when evaluated on different test collections for the two encoding schemes. Bold values highlight the best results per collection and metric.

Overall, the $\text{GPT}^{\texttt{multi}}_{\text{FC}}$ method had higher results than the $\text{GPT}^{\texttt{multi}}_{\text{CNN}}$ method on the CIPI and PS datasets, especially on Acc_0 and Acc_1. For CIPI, $\text{GPT}^{\texttt{multi}}_{\text{FC}}$ surpassed $\text{GPT}^{\texttt{multi}}_{\text{CNN}}$ with gains of 5.4% in Acc_0, 0.6% in Acc_1, and 0.1 in MSE. For PS, $\text{GPT}^{\texttt{multi}}_{\text{FC}}$ slightly exceeded $\text{GPT}^{\texttt{multi}}_{\text{CNN}}$ with a 3.7% improvement in Acc_1 and a 0.6-point reduction in MSE, while Acc_0 was nearly equal for both methods, although $\text{GPT}^{\texttt{multi}}_{\text{CNN}}$ had a smaller standard deviation. Both methods displayed similar performance on the FS dataset, with less than a 1% difference across all metrics. As a result, subsequent experiments will reference the $\text{GPT}^{\texttt{multi}}_{\text{FC}}$ model.

The comparison between Tables 2 and 3 shows a trend change with better results performed with the FC version of the models. The other major difference is the relative improvement between the $\text{GPT}^{\texttt{multi}}_{\text{FC}}$ method and the best previous model GPT_{CNN} in the CIPI and slightly in the PS dataset. In contrast, the FS dataset results remain comparable. In CIPI, Acc_0 is 11.3% higher in $\text{GPT}^{\texttt{multi}}_{\text{FC}}$, and in PS, there is a relative improvement of 12.8%. For CIPI, Acc_1 sees a minor increase of 0.4%. MSE exhibits a small improvement of 3.6% for CIPI and 0.5% for PS. Possible reasons include label quality differences—CIPI annotated by a musicology team, PS labels provided by the platform, and FS crowdsourced by users—or the impact of dataset sizes—CIPI being the smallest and FS the largest.

5.3 Ranking generalization experiment

In this experiment, we assess the ranking capabilities of the proposal in a zero-shot setting by utilizing the embeddings of the projection layer of the model (check Figure 3). We reduce the 768-dimensional embeddings to a single dimension using Principal Component Analysis (PCA) and employ the resulting values to rank the target pieces.

Table 4 shows the results obtained resorting to the

Kendall rank correlation coefficient, τ_c, for all data collections discussed in the experiment, considering both the single-task and multi-task frameworks posed. Note that MK and HV are only used for benchmarking purposes.

Train	Evaluation				
	CIPI	PS	FS	MK	HV
CIPI	.67 (.01)	.56 (.02)	.56 (.01)	.67 (.05)	.50 (.05)
PS	.67 (.01)	.58 (.02)	.56 (.01)	.68 (.01)	.43 (.04)
FS	.64 (.04)	.55 (.01)	.56 (.02)	**.71 (.02)**	**.56 (.07)**
MULTI	**.68 (.02)**	**.59 (.02)**	**.56 (.01)**	.63 (.02)	.51 (.07)

Table 4. Zero-shot ranking results. Bold values denote the best-performing result on each evaluation dataset.

In the three training datasets, the multi-task architecture $\text{GPT}_{\text{FC}}^{\text{multi}}$ achieves the best performance with CIPI ($\tau_c = 0.68$), PS ($\tau_c = 0.59$), and FS ($\tau_c = 0.56$). Unexpectedly, the FS method outperforms others in the datasets of the MK ($\tau_c = 0.61$) and HV ($\tau_c = 0.56$). This outcome may suggest that simultaneous training on all three datasets could limit generalizability. Alternatively, the presence of license-free pieces composed after 1900 in the FS dataset, which users have uploaded, might explain the difference.

The HV dataset displays notably lower generalizability, possibly due to the smaller number of pieces, resulting in higher standard deviations. Potential bias similar to MK could also arise from the predominance of pre-20th-century data in CIPI and PS. These factors might affect the zero-shot experiment's performance. However, we must also acknowledge that most composers used for training are white males, and the HV results are significantly worse than the rest of the datasets. Therefore, future research should investigate and minimize the potential gender gap in difficulty prediction tasks.

5.4 Comparison with previous approaches

This last experiment compares the goodness of the proposed methodology in sheet music scores against other image-based approaches and with the symbolic-oriented methods domain. Regarding sheet image methods, we consider the reference method by Tsai et al. [23] based on bootleg mid-representation, denoted as GPT_{EMB}. Concerning the symbolic baseline, we reproduce the approach in [4] that proposes to describe the symbolic score in terms of piano fingering information, expressive annotations, and pitch descriptors to feed a recurrent model based on Gated Recurrent Units with attention layers (referred to as GRU+Att). Table 5 provides the results obtained. For comparative purposes, we only consider the CIPI dataset as the reference symbolic work accounted for that collection.

Examining the experiments, the $\text{GPT}_{\text{FC}}^{\text{multi}}$ model may be observed to outperform the other cases in the Acc_0 figure of merit. However, for the rest of the metrics, the reference symbolic case—denoted as GRU+Att—outperforms all image-oriented recognition models. Such a fact suggests that, while a bootleg score somehow suits this difficulty estimation task, a performance gap between this representation and pure symbolic notation needs to be addressed.

Case	Acc_0 (%)	Acc_1 (%)	MSE
Symbolic [4]			
GRU+Att	39.5(3.4)	**87.3(2.2)**	**1.1(0.2)**
Tsat et al. [23]			
GPT_{EMB}	19.7(4.0)	58.1(7.2)	3.3(0.8)
Proposal			
GPT_{FC}	34.3(6.1)	78.1(4.6)	1.6(0.3)
GPT_{CNN}	36.2(8.2)	81.7(1.5)	1.4(0.1)
$\text{GPT}_{\text{FC}}^{\text{multi}}$	**40.3(4.3)**	82.0(1.4)	1.3(0.1)

Table 5. Performance results for the symbolic [4] and Tsai et al. [23] methods as well as the proposed approach for the CIPI dataset. Bold values highlight the best result per figure of merit.

Finally, the GPT_{EMB} model achieves the lowest performance of all alternatives, with remarkably lower accuracy rates than our proposal. Note that such a fact emphasizes the relevance of our work as a more suitable approach for performing difficulty estimation in sheet music images.

6. CONCLUSIONS

Estimating the performance difficulty of a music piece is a crucial need in music education to structure the learning curriculum of the students adequately. This task has recently gathered attention in the Music Information Retrieval field, given the scarce existing research works devoted to symbolic machine-readable scores. However, due to the limited availability of this type of data, there is a need to devise methods capable of addressing this task with image-based sheet music.

Attending to its success in related classification tasks, this work considers the use of a mid-level representation—namely, bootleg score—that encodes the content of a sheet music image with a GPT-based recognition framework for predicting the difficulty of the piece. Instead of directly applying this methodology, we propose using specific embedding mechanisms and multi-task learning to reduce the task complexity and improve its recognition capabilities. The results obtained with five different data collections—three of them specifically compiled for this work—prove the validity of the proposal as it yields recognition rates comparable to those attained in symbolic machine-readable scores.

Further work comprises assessing and proposing alternative representations to the bootleg scores (*e.g.*, solutions based on Optical Music Recognition). Also, we consider that using smaller training sequences using hierarchical attention models or weak labels for varying-length piece fragments may report benefits in the process. Finally, the practical deployment of this proposal in real-world scenarios involving real users may report some additional insights about the validity of the proposal.

7. ACKNOWLEDGMENT

We want to thank T.J. Tsai and all his students, especially Daniel Yang, for having conducted the prior research on the bootleg score and, above all, for sharing all their work in the interest of Open Science. We are also grateful to Pedro D'Avila for bringing to our attention the work of Alejandro Cremaschi related to the Hidden Voices project. Lastly, we thank Alejandro Cremaschi and the University of Colorado Boulder Libraries team, David M. Hays and Jessica Quah, for providing us with the scores.

This work is funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI) within the Musical AI Project – PID2019-111403GB-I00/AEI/10.13039/501100011033 and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) (NRF-2022R1F1A1074566).

8. REFERENCES

[1] V. Sébastien, H. Ralambondrainy, O. Sébastien, and N. Conruyt, "Score analyzer: Automatically determining scores difficulty level for instrumental e-learning," in *Proceedings of 13th International Society for Music Information Retrieval Conference, ISMIR*, Porto, Portugal, 2012.

[2] S.-C. Chiu and M.-S. Chen, "A study on difficulty level recognition of piano sheet music," in *Proceedings of the IEEE International Symposium on Multimedia*. IEEE, 2012, pp. 17–23.

[3] E. Nakamura and K. Yoshii, "Statistical piano reduction controlling performance difficulty," *APSIPA Transactions on Signal and Information Processing*, vol. 7, 2018.

[4] P. Ramoneda, D. Jeong, V. Eremenko, N. C. Tamer, M. Miron, and X. Serra, "Combining piano performance dimensions for score difficulty classification," *arXiv preprint arXiv:2306.08480*, 2023.

[5] "Musescore have automatic difficulty categories from year 2022," https://musescore.com/, accessed on April 11, 2023.

[6] "Ultimate guitar have automatic difficulty categories from year 2022," https://www.ultimate-guitar.com/, accessed on April 11, 2023.

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

[8] E. Nakamura, N. Ono, and S. Sagayama, "Merged-output hmm for piano fingering of both hands." in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*, Taipei, Taiwan, 2014, pp. 531–536.

[9] E. Nakamura and S. Sagayama, "Automatic piano reduction from ensemble scores based on merged-output hidden markov model," in *Proceedings of the 41st International Computer Music Conference, ICMC*, Denton, USA, 2015.

[10] P. Ramoneda, N. C. Tamer, V. Eremenko, M. Miron, and X. Serra, "Score difficulty analysis for piano performance education," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Singapore, Singapore, 2022, pp. 201–205.

[11] 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*, 2019, pp. 908–915.

[12] J. Cheng, Z. Wang, and G. Pollastri, "A neural network approach to ordinal regression," in *Proceedings of the IEEE International Joint Conference on Neural Networks, IJCNN*. Hong Kong, China: IEEE, 2008, pp. 1279–1284.

[13] D. Bennett, S. Macarthur, C. Hope, T. Goh, and S. Hennekam, "Creating a career as a woman composer: Implications for music in higher education," *British Journal of Music Education*, vol. 35, no. 3, pp. 237–253, 2018.

[14] J. Halstead, *The woman composer: Creativity and the gendered politics of musical composition*. Routledge, 2017.

[15] R. Cutietta, "Content for music teacher education in this century," *Arts Education Policy Review*, vol. 108, no. 6, pp. 11–18, 2007.

[16] J. Magrath, *Pianists guide to standard teaching and performance literature*. Alfred Music, 1995.

[17] J. Calvo-Zaragoza, J. H. Jr, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–35, 2020.

[18] D. Yang, T. Tanprasert, T. Jenrungrot, M. Shan, and T. Tsai, "MIDI passage retrieval using cell phone pictures of sheet music," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR*, Delft, The Netherlands, 2019, pp. 916–923.

[19] D. Yang, A. Goutam, K. Ji, and T. J. Tsai, "Large-scale multimodal piano music identification using marketplace fingerprinting," *Algorithms*, vol. 15, no. 5, p. 146, 2022.

[20] D. Yang, K. Ji, and T. Tsai, "Aligning unsynchronized part recordings to a full mix using iterative subtractive alignment," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR*, Online, 2021, pp. 810–817.

[21] K. Ji, D. Yang, and T. Tsai, "Piano sheet music identification using marketplace fingerprinting," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR*, Online, 2021, pp. 326–333.

[22] D. Yang and T. J. Tsai, "Piano sheet music identification using dynamic n-gram fingerprinting," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, pp. 42–51, 2021.

[23] T. Tsai and K. Ji, "Composer style classification of piano sheet music images using language model pretraining," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, Montreal, Canada, 2020, pp. 176–183.

[24] D. Yang and T. Tsai, "Composer classification with cross-modal transfer learning and musically-informed augmentation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR*, Online, 2021, pp. 802–809.

[25] K. Ji, D. Yang, and T. J. Tsai, "Instrument classification of solo sheet music images," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Toronto, ON, Canada, 2021, pp. 546–550.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[27] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in *Advances in Neural Information Processing Systems*, 2022.

[28] "Piano street," https://www.pianostreet.com/, accessed on April 11, 2023.

[29] "Free-scores," https://www.free-scores.com/, accessed on April 11, 2023.

[30] University of Colorado, "Hidden voices project," https://www.colorado.edu/project/hidden-voices/, accessed on April 11, 2023.

[31] H. Walker-Hill, *Piano Music by Black Women Composers: A Catalog of Solo and Ensemble Works*, ser. Music Reference Collection. Greenwood Press, 1992.

[32] L. Gaudette and N. Japkowicz, "Evaluation methods for ordinal classification," in *Proceedings of the 22nd Canadian Conference on Advances in Artificial Intelligence*, Kelowna, Canada, 2009, pp. 207–210.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.