Modelling Flute Difficulty through a Corpus of Annotated Scores

RÉSUMÉ

Modeling musical instrument difficulty helps us understand and model the challenges performers face when learning or interpreting music. We present an open and curated corpus of 69 public domain flute pieces to analyze flute performance difficulty, including repertoire and extracts from Wilhelm Popp's 19th-century flute method. The pieces are annotated by professional flute educators working with the grading system of the Associated Board of the Royal Schools of Music (ABRSM). We study inter-annotator agreement and discuss key pedagogical dimensions of flute performance difficulty. The corpus analysis reveals relationships between musical attributes and difficulty levels. We finally introduce and evaluate a kNN baseline model for predicting flute performance difficulty using symbolic music features. This study helps in understanding elements for assessing instrumental difficulty.

Modéliser la difficulté du jeu instrumental aide à comprendre et à analyser ls défis auxquels sont confrontés les interprètes sont confrontés lorsqu'ils apprennent ou jouent une œuvre. Nous présentons un corpus ouvert qui rassemble 69 pièces dans le domaine public pour flûte traversière solo. Ce corpus regroupe des pièces du répertoire tout comme des extraits de la méthode pour flûte du XIXsiècle de Wilhelm Pop. Les pièces sont annotées par des professeur·es de flûte, en utilisant le système de classification du Associated Board of the Royal Schools of Music *(ABRSM). Nous étudions la concordance entre les annotateurs et discutons de quelques facteurs contribuant à cette difficulté d'interprétation, qui sont en partie confirmés par l'analyse statistique du corpus. Enfin, nous introduisons et évaluons un modèle simple kNN (k plus proches voisins) pour prédire la difficulté d'exécution à la flûte traversière à partir de ces facteurs. Cette étude contribue à mieux comprendre les éléments contribuant à la difficulté instrumentale.*

1. INTRODUCTION

The musical difficulty perceived by a musician when interpreting a piece of music is related to multiple aspects of the score and of the instrument characteristics, such as pitch range, rhythmic complexity, and technical demands. Assessing the difficulty of a musical score is crucial in various pedagogical and performance contexts. In instrumental learning, it supports a coherent progression of skill acquisition by aligning repertoire with student ability levels [22]. It is also essential in music education institutions when selecting examination repertoire [1, 4, 2, 3], and in ensemble settings to ensure parts are appropriately matched to performers' technical proficiency [12].

1.1. Models and Datasets for Instrument Difficulty

Several studies have proposed methods to assess musical difficulty. For example, Mazur and Łaguna [29] emphasise the importance of defining achievement levels, based on technical ability, expressiveness, and overall impression. Andono et al. [7] introduced a method for melody difficulty classification using frequent patterns and internote distance analysis.

Prior research has explored the automatic modelling of instrument-specific difficulty referencing existing music scores, with a predominant focus on piano and, to a lesser extent, guitar, saxophone and violin.

For *piano*, the Score Analyzer system automates difficulty classification identifying key features such as tempo and hand displacement [44]. Chiu and Chen [14] proposed a regression-based system using symbolic features such as pitch entropy and fingering complexity, trained on MIDI datasets with annotated difficulty levels. Ramoneda et al. introduced an HMM-based model to infer fine-grained fingering patterns from symbolic data, which correlate with both local and global difficulty levels [41]. A hybrid deep learning pipeline was developed combining piano roll inputs and handcrafted features for piano score classification [19]. Expressive modelling was also incorporated through embeddings from systems such as VirtuosoNet, which capture phrasing, dynamics, and other expressive elements [38]. RubricNet is based on rubric-style descriptors that reflect common pedagogical dimensions, such as pitch entropy, displacement rate, and rhythmic density [37]. Efforts to model piano arrangement difficulty, such as the work by Gover and Zewi [20], have focused on generating arrangements at varying playing levels. Their approach uses a parallel dataset of annotated arrangements and deep learning models to translate between difficulty levels while maintaining the musical integrity of the piece. Recently, Ramoneda et al. advanced difficulty-controlled simplification using synthetic data, employing a transformer-based method for MusicXML scores that preserves melody and harmony while adjusting difficulty levels [40].

Published piano corpora for supervised difficulty classification with symbolic and image-based models include Mikrokosmos-difficulty (147 pieces by Béla Bartók) [41], Can I Play It? (652 piano pieces, with public domain scores and 9 difficulty levels according to Henle Verlag labels) [38], Hidden Voices (17 pieces by Black female

composers) [42], and PianoPairs (3,000 pieces with 128 synthetic variations each, at different difficulty) [40].

For *guitar*, Tandon and Tandon [45] use user-specific labels and active learning to predict difficulty from segment-level tab features. Velezvasquez et al. [48] present a rule-based baseline relying on chord annotations and timings. Pedagogical recommendation systems were proposed by Hassein-Bey et al. [24] (chord, rhythm and tempo features) and Müllerschön et al. [31] (tablature profiling with educational tags and psychomotor difficulty). For *saxophone*, Libřický and Hajič jr [28] introduced a cost-of-traversal approach, modeling difficulty as transitions between fingering pairs through physical/biomechanical constraints. They publicly release a difficulty dataset of 817 recorded trills covering 741 fingering pairs with extracted trill speeds. In a broader multi-instrumental study, Deconto et al. [18] investigated automatic difficulty classification for *violin*, piano, and acoustic guitar using features extracted from scores, training classifiers on transcribed pedagogical repertoire.

However, very few studies address other instruments, and we are not aware of any MIR datasets on modelling difficulty for flute. Several computational studies on flute playing focus on technical elements such as breath control (embouchure and airflow) and fingering accuracy [16, 23], while none specifically address the computational modelling of flute difficulty.

1.2. Contents

This paper extends previous research on instrument difficulty estimation by focusing on the *flute*. We briefly present syllabi from grading systems across different countries (Section 2) and review elements contributing to difficulty (Section 3). We propose a new dataset of 69 public domain flute pieces, including extracts from a flute learning method by Wilhelm Popp (1829-1902) and selections of repertoire pieces. Expert flute teachers provided annotations in the Associated Board of the Royal Schools of Music (ABRSM) system (Section 4). We study inter-annotator agreement, and, when there is a reasonable agreement, we define the *reference grade* as the median of these annotations grades provided by the teachers. We explore statistics on this dataset, proposing metrics to assess difficulty. A simple classifier based on some of these metrics achieves a 78.0% accuracy in predicting difficulty with a tolerance of one grade (Section 5). We conclude by discussing how such datasets may help in modelling difficulty, with a view towards future applications in difficulty-aware arrangement and ensemble adaptation (Section 6).

2. GRADING SYSTEMS AND SYLLABI

A *syllabus* in music education outlines theoretical knowledge, technical skills, and musicianship to be assessed at various levels, serving both as a *teaching guide* and a *benchmark for exams*, allowing for consistency across different regions or countries [47]. *Syllabi*-based datasets are thus already being leveraged in MIR research to align computational models with educational grading standards [18, 39]. When applied to instrumental practice, these syllabi often include a required repertoire, encompassing both traditional/classical and modern pieces, which may change annually or at longer intervals. This repertoire serves both pedagogical and evaluative purposes, aligning technical and expressive goals with graded difficulty levels, measuring *personal progress* and *musical attainment* [47].

Music learning systems vary widely in different countries, even at the beginner or amateur level [50]. In some contexts, a single instrument teacher may provide both instrumental instruction and music theory, while in others, students may work with two to four different instructors – each responsible for areas such as general music education, choir singing, or other complementary training. Beyond grades, each syllabus offers distinct approaches to music learning. For example, the Central Conservatory of Music (CCOM) in China [2] emphasises technical proficiency through a structured grading system for both Western classical instruments and traditional Chinese instruments, requiring mastery of core techniques and memorisation of repertoire from early stages. Some widely used systems, even outside their home countries, are the ones from the Royal Conservatory of Music (Canada) [3], and, from the UK, the Associated Board of the Royal Schools of Music (ABRSM) [1] and Trinity College [4].

We focus in this study on ABRSM. Founded in 1889, the ABRSM is one of the oldest and most widely adopted music examination boards worldwide. It provides graded instrumental and theory syllabi from beginner to advanced grades, with a strong emphasis on structured technical progression, sight-reading, scales, and repertoire drawn primarily from the Western classical tradition. Nowadays, ABRSM examinations are taken in over 90 countries, making its grading system a common international reference point for music education, pedagogy, and curriculum design [1]. In the remainder of the paper, we focus on grades up to Grade 6 in the ABRSM system, as these levels are commonly encountered in music education and are often the focus of pedagogical studies.

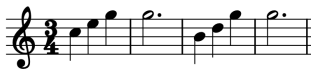
3. WHAT MAKES FLUTE PLAYING DIFFICULT?

The difficulty of flute performance is multifaceted, involving technical, physiological, and cognitive challenges that vary across different levels and age groups [36, 10, 43, 25, 26].

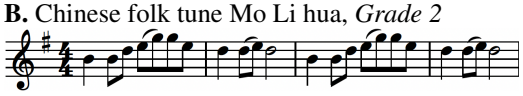
Range, rhythmic and key signature complexity steadily increase from beginner (Reference Grade 1–2) to intermediate levels (Grades 5–6) [5]. For example, flowing semiquaver runs and subtle syncopations require precise rhythmic control (Figure 1C). *Long notes and scale passages* also present significant challenges, reinforcing the importance of breath control, tone consistency and fingering across registers.

Performing larger leaps not only challenges flexibility

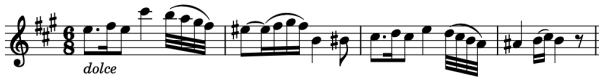
A. Popp, *Grade Initial (0)*



B. Chinese folk tune Mo Li hua, *Grade 2*



C. Popp, piece No.4, Larghetto, *Grade 5*



D. Philippe Gaubert, Madrigal (1905), *Grade 5.5*

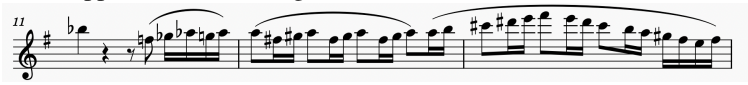


Figure 1: The corpus contains exercises from Popp method (A, C) and other public domain pieces (B, D). Reference grades are defined in Section 5.1.

but also places greater demands on posture stability and breath control [49]. Meanwhile, *interval jumps*, even between adjacent notes like C–D, are technically demanding because they involve substantial fingering rebalancing and embouchure adjustments [26]. Small intervals with accidentals are more difficult (e.g. Figure 1D).

Breath management is a particularly critical dimension for flutists. Compared to other woodwinds, the flute demands the greatest amount of airflow with the least resistance, making breath control fundamentally more difficult [21]. Beginners often struggle with airflow stability, leading to issues in onset, register shifts, and sustaining long phrases [15]. For example, holding the sustained G in Figure 1A is already challenging for beginners. Bulut [13] further highlights that young learners, due to smaller lung capacity, require adaptive phrasing strategies. *Tone quality* is a key aspect of flute playing, and is also linked to breath-centered technique and musical sensitivity [8].

As with other woodwind instruments, *articulation* is a crucial factor in developing speed and precision. Effective articulation involves coordinating the tongue, air support, and embouchure to produce clean, rapid note transitions. The need to master single and double tonguing appears more prominently after Grade 4 [5], aligning with syllabus demands and confirmed by pedagogical resources [17, 46]. Clear articulation under dynamic constraints (*pp* to *ff*) becomes critical as students advance.

4. GATHERING AND ANNOTATING A CORPUS FOR FLUTE DIFFICULTY

4.1. Scores

The corpus aims to provide a broad representation of pieces used in formal music education within Initial grade (Grade 0) to Grade 7, while ensuring that all works are in the public domain. It contains:

- 30 pieces selected from the original edition (1887) of Wilhelm Popp's *Erster Flöten-Unterricht* method [34]. Popp (1828–1903) was a key figure in flute pedagogy, and his method book remains a valuable source for understanding flute technique and musical education of the time (Figures 1A, 1C, 2). We selected short excerpts (fewer than six lines), focusing on balanced phrases with limited local variation.
- 39 repertoire pieces referenced with corresponding grades in three well-known music education syllabi: ABRSM [5], Trinity [6], and RCM [33] (Figures 1B, 1D).

These pieces collectively represent a variety of difficulty levels and pedagogical focuses used in formal flute instruction. Selections were made based on the availability of public domain scores [1] and the representativeness of each piece in relation to graded difficulty. As expected, since this corpus is limited to public domain works, there are very few 20th-century pieces. Nevertheless, the selected repertoire exhibits a certain stylistic and technical diversity, with a range of musical forms, articulations, and expressive demands.

4.2. Data Collection

The Wilhelm Popp's method book for the flute has been digitized by the Boston Public Library [35]. The repertoire pieces were sourced from public domain sites, mostly from IMSLP (www.imslp.org). Some scores were encoded using the Musescore importer [30], available to registered users, which relies on Audiveris OMR [9]. All scores were manually edited, corrected, particularly to encode musical elements such as enharmonics, articulation, dynamics, ornaments, and phrasing, to reflect the source images. Note that, for the repertoire pieces, we did not undertake critical editorial work, which is beyond the scope of this study.

4.3. Grading

Three professional flute teachers, co-authors of this article, evaluated the difficulty level of the pieces of the dataset. Those teachers respectively have 5, 7, and 30 years of experience teaching within the ABRSM system, and hold degrees in flute performance and flute pedagogy. They put their own grade in ABRSM system, on each of the 69 pieces, based on their expertise and knowledge of the syllabi requirements. A level annotation corresponds to the level required to complete a grade. When a teacher felt a piece falls between two grades, or when they hesitate, they used half values such as 3.5 for "between Grade

[1] Public domain under EU law grants protection for 70 years after the composer's death (with up to 14 additional years of war extensions applicable in France)

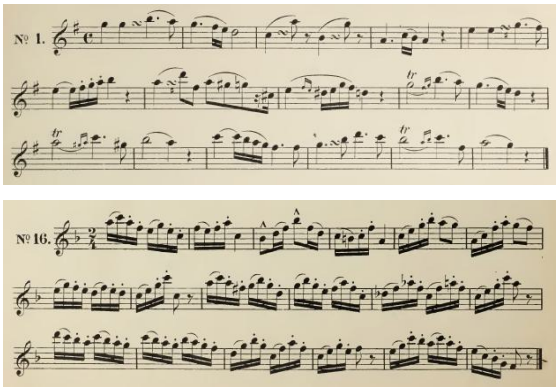


Figure 2: Scores with disagreement among teachers. a) Popp, p. 25 n.1, reported grades from 4 to 6.5. b) Popp, p. 22 n.16, reported grades from 2 to 4.5.

3 and 4". Of course, student progression is not strictly linear and grading can sometimes be subjective.

5. RESULTS

In this section, we study inter-annotator agreement (5.1), and then how musical parameters correspond to their grading. We begin by analysing basic statistics related to key, pitch, intervals, and time signature (5.2) and then define and examine four additional features designed to model musical difficulty (5.3). We test a simple kNN classifier on those features (5.4).

5.1. Inter-annotator Agreement and Reference Grades

The Krippendorff's α for this annotation task, computed here on all pairs of teachers on every graded piece, is here $\alpha = 0.855$ indicating a high level of agreement among the three teachers [27, 32]. More precisely, on the 69 pieces, 8 (11.6%) have the exact same grading across the three teachers, 47 (68.1%) within 1 grade, and 59 (85.5%) within 1.5 grade. Finally, disagreement spanning more than two grades is shown in 10 pieces, as for example:

- (Figure 2a.) One of the teachers emphasised its expressive nature, noting that varied articulation, dense ornamentation, and nuanced phrasing require technical control and careful breath management (grade 6.5). Whereas the other teachers considered the overall difficulty to be more moderate (grades 4 to 4.5), both agree that the frequent turns, trills, and grace notes constitute the primary source of difficulty, even though the piece is relatively short and the underlying melody remains clear.
- (Figure 2b). One of the teachers described it as relatively simple (grade 2), highlighting its rhythmic and articulatory homogeneity, straightforward phrasing, and comfortable range despite the presence of running semiquavers. Whereas the other teachers emphasised the technical demands of continuous semiquaver patterns, tonguing coordination, and stable breath support in the middle register, both agree that the piece is structurally clear but places sustained demands on execution in practice (grades 4 to 4.5).

These differences highlight the subjective perceived nature of difficulty assessment, where expressive demands, technical consistency, and pedagogical priorities can lead to divergent yet equally valid evaluations.

In the following, we exclude these 10 pieces and, for each of the 59 remaining pieces, we consider the reference grade as the *median* of the grades provided by the three teachers. Note that 13 of these 59 pieces were also referenced in at least one ABRSM syllabus. On all these pieces, except one, the reference grading computed here is at most 0.5 apart from this official ABRSM grading. Altogether, these 59 pieces are distributed across reference grades. Each half-grade is represented by between 2 and 8 pieces, and, grouping half-grades with their nearest lower full grade, each full grade is represented by between 5 and 12 pieces.

5.2. Key, Pitches, Intervals, and Time Signature

In Figures 4 to 3, half-grades are grouped with their nearest lower full grade. Lower-grade pieces favor simple *key signatures*: About 67.8% of the pieces fall within keys of no more than one sharp or flat (Figure 5). Relative minors are generally less represented and appear in slightly more advanced levels. *Pitch range* expands across grade levels, with higher-grade pieces including lower (e.g., C4 or even B3) and higher pitches (E6–A6) (Figure 3). At higher levels, pieces exhibit a wider range of *intervals*, including above the octave (Figure 4). However, all pieces usually include many seconds (M2, m2), such as some examples on Figure 1. Common *time signatures* such as 4/4, 3/4, and 2/4 dominate the dataset. Other time signatures are mostly associated with higher levels (Figure 6).

5.3. Note Density, Accidentals, Fingerings, Intonation

Note density is measured as the average number of notes per measure. Both Popp and syllabus pieces show a clear trend toward higher note density at more advanced grade levels (Figure 7). Lower levels generally cluster around 3–6 notes per measure, while upper levels exceed 10 notes in many cases.

Figure 7b shows the distribution of accidental difficulty (including key-signature accidentals) across grade levels. For each (diatonic) pitch, we define the *accidental difficulty* by the diversity of accidentals per pitch class in a given piece, with a maximum value of 2 indicating that a single pitch appears in natural, sharp, and flat forms, combined with the proportion of notes carrying an accidental in a piece. The average accidental difficulty generally increases with grade level.

Average fingering changes per note reflect the number of fingers moving between each pair of pitches (Figure 7c), considering common fingerings [11]. Such a met-

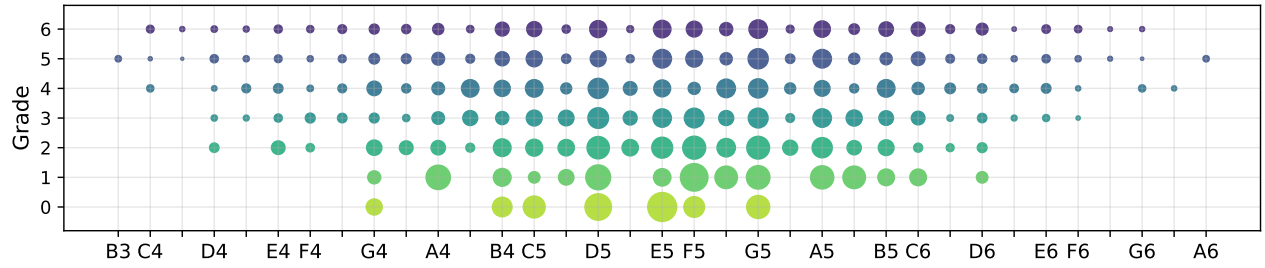


Figure 3: Average per-piece pitch distribution by reference grade. Note that the small number of pieces on each grade can bring bias (as here, no F5 on the 10 pieces of Grade 2–2.5).

piece	ref. grade	key	pitch range	time signature	interval	density	accidentals	fingering	intonation
A. Popp	0	C major [2]	g' →g" [2]	3/4 [1]	2.88	2.89	0	2.88	5.40
B. Mo Li hua	2	G major [1]	d' →g" [3]	4/4 [1]	2.18	4.80	0	2.34	6.23
C. Popp	5	A major [5]	a' →f"'[5]	6/8 [4]	2.84	7.74	0.39	2.80	7.66
D. Madrigal	5.5	G major [1]	c'→g"' [5]	4/4 [1]	2.07	7.71	0.55	2.81	7.76

Table 1: Difficulty metrics for example pieces (see Figures 1 and 7). These metrics are computed for the whole pieces, not only the excerpts. The values between square brackets show, according to ABRSM guidelines for *sight reading*, the minimal grade for playing such key/pitch range/time signatures.

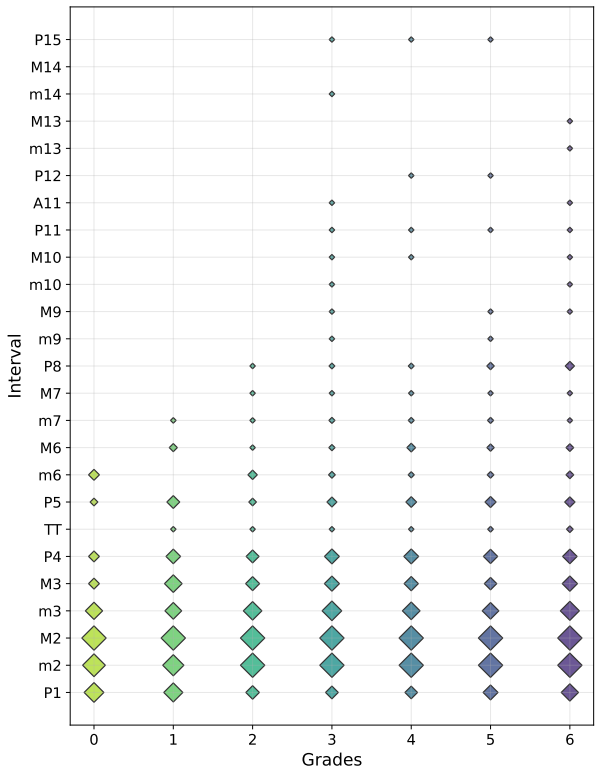


Figure 4: Average per-piece interval distribution by grade. Intervals are counted in semitones; for example, minor seconds (m2) also include occasional chromaticisms (augmented unisons, A1).

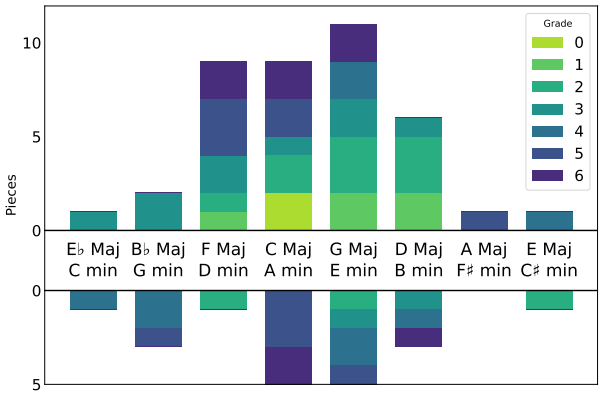


Figure 5: Key distribution by grade.

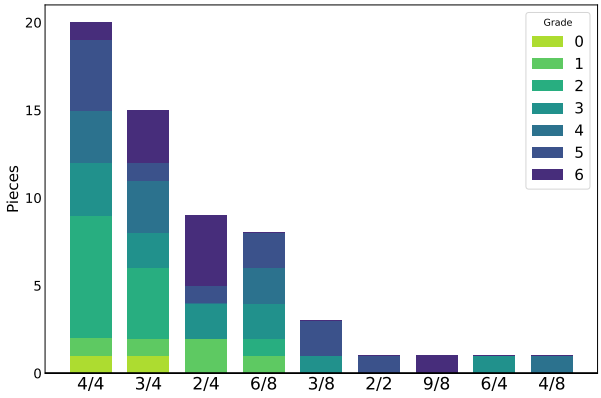


Figure 6: Time signature distribution by grade.

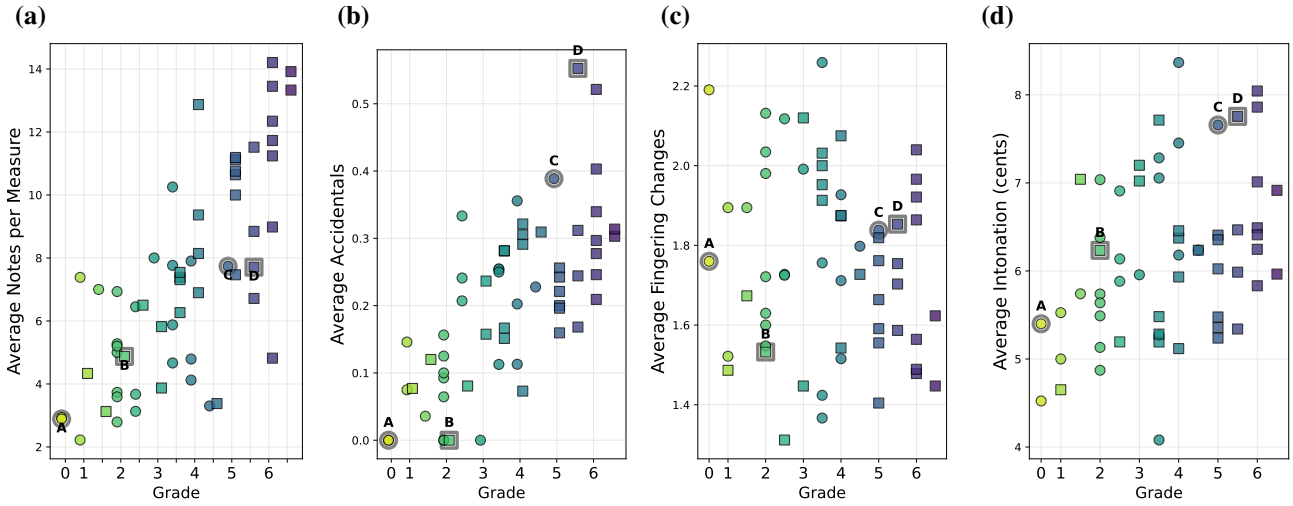


Figure 7: Piece distribution by grade according to (a) note density, (b) average accidentals, (c) average fingering changes, and (d) average intonation. Popp pieces: circles. Syllabus pieces: squares. The A/B/C/D marks refer to the pieces shown on Figure 1 and Table 1. A small horizontal offset separates Popp pieces and syllabus pieces.

ric does not consistently increase with grade. Further research could include better modeling of the physical effort calculation for such transitions, beyond a simple number of fingers. For example, contrary movements of the different fingers are harder and may impact the instrument balance.

Both the Popp and syllabus collections also exhibit a wide range of *intonation difficulty* across grades (Figure 7d). This metric reflects how far the sound produced without precise mouth control may deviate from ideal tuning, with higher values indicating notes that are harder to play in tune on the flute [11]. While precise intonation is an essential skill that improves over time, limited intonation accuracy does not prevent students from performing simpler pieces, as for example, the *Mo Li hua* piece that requires comparatively high pitch control (Table 1).

5.4. Modelling Difficulty

The difficulty of a piece usually does not arise from *all* metrics simultaneously. For example, as shown in Table 1, the *Madrigal* (D) is based on the relatively simple key of G major and involves few fingering changes. However, it exhibits advanced characteristics, including a high note density (7.71 notes per measure), numerous modulations and accidentals.

We tested on the 59 pieces a simple kNN classifier ($k = 5$) using a *leave-one-out* cross-validation strategy, where each model was trained on 58 pieces and tested on the remaining one. The classifier relies on six features that are related to instrumental difficulty: note density, interval, accidentals, pitch-range, key, and time signatures (Figure 2 to Figure 6). The model's predictions reasonably align with the reference dataset: the overall accuracy is 28.8%, and reaches up to 78.0% when allowing a tolerance of ± 1 grade, suggesting that the selected features capture meaningful aspects of perceived instrumental difficulty. More sophisticated models could likely yield improved performance. Naturally, models dealing with full score data would also enhance predictive accuracy, but these initial results already demonstrate that the selected features provide a solid foundation for modelling musical difficulty.

6. DISCUSSION, DATA RELEASE, AND PERSPECTIVES

Difficulty on the flute – as on other instruments – arises from multiple factors, motivating representations that can support flexible pedagogical contexts. This study introduces a new dataset of 69 annotated pieces, including 59 with reference grades, and proposes several metrics to evaluate difficulty, some of which show meaningful correlation with reference difficulty levels. We release scores of the 69 pieces along with the annotations made by the three teachers on these 69 pieces, under the open CC-BY-SA-4.0 license, on the git repository available at [REDACTED]. Beyond difficulty estimation itself, this model can serve as a building block for difficulty-aware arrangement and ensemble adaptation.

This study has several limitations and potential avenues for future research. The dataset creation process may involve certain biases. Expanding the dataset with more diverse sources and teaching methods would enhance robustness. Including more beginner-level pieces would provide a more balanced representation across all skill levels.

The feature set could be improved with elements such as articulation patterns and phrasing complexity. A better model could account for biomechanical aspects, such as finger coordination and force, as well as alternative fingerings used at more advanced levels, and the modeling of the difficulty of each interval. Note density does not account for nuanced aspects such as varied durations, the challenges of both long and short notes, and tuplets. Including these detailed rhythmic elements could refine the difficulty assessment. Additionally, dynamic changes

and articulations were not modelled, though they significantly influence technical and interpretative complexity. Integrating these features in a joint framework may better capture the actual physical and technical effort required in performance.

Altogether, this study lays the foundation for modelling instrumental difficulty, and opens the door to further research on the flute and other instruments to better understand and support music pedagogy.

7. REFERENCES

[1] Associated board of the royal schools of music. https://www.abrsm.org/en-gb. Accessed: 2026-03-10.

[2] Flute examination syllabus. Central Conservatory of Music (CCOM), https://www.kaoji.com/#/information?informationType=2&activeName=two-request. Accessed: 2026-03-10.

[3] The royal conservatory of music. https://www.rcmusic.com/. Accessed: 2026-03-10.

[4] Trinity college london. https://www.trinitycollege.com/. Accessed: 2026-03-10.

[5] Abrsm woodwind syllabus 2022 onwards (includes flute, clarinet, oboe, bassoon, and saxophone). Associated Board of the Royal Schools of Music; https://www.abrsm.org/en-gb/instruments/woodwind, 2022. Accessed: 2026-03-10.

[6] Woodwind 2022 practical syllabus. Trinity College London, https://www.trinitycollege.com, 2023.

[7] Pulung Nurtantio Andono, Edi Noersasongko, Guruh Fajar Shidik, Khafiizh Hastuti, Sudaryanto Sudaryanto, and Arry Maulana Syarif. Melody difficulty classification using frequent pattern and internotes distance analysis. *International Journal of Advanced Computer Science and Applications*, 13(2), 2022.

[8] Jessica Benevento. The french flute school: A flute curriculum. *The Flutist Quarterly*, 46(4):46–51, 2021.

[9] Hervé Bitteur and contributors. Audiveris: Open-source optical music recognition, 2025. Accessed: 2026-03-10.

[10] Janice Dockendorff Boland. *Method for the One-Keyed Flute: Baroque and Classical*. University of California Press, 1998. Winner of the National Flute Association's Newly Published Music Competition, 1999.

[11] Andrew Botros. The virtual flute. https://flute.fingerings.info/, 2001–2014. Accessed: 2026-03-10.

[12] Richard Brooks. The importance of arrangements for amateur ensembles, 2022. Accessed: 2026-03-10.

[13] Seyhan Bulut. A comparison of the first three flute lessons with beginner and intermediate-advanced level flute students. In *Procedia - Social and Behavioral Sciences*, volume 177, pages 229–234, 2015.

[14] Shih-Chuan Chiu and Min-Syan Chen. A study on difficulty level recognition of piano sheet music. In *Proceedings of the 2012 IEEE International Symposium on Multimedia (ISM)*, pages 17–23. IEEE, 2012.

[15] Patricio De La Cuadra, Benoît Fabre, Nicolas Montgermont, and Christopher Chafe. Analysis of flute control parameters: A comparison between a novice and an experienced flautist. *Acta Acustica united with Acustica*, 94(5):740–749, 2008.

[16] Andrey R. da Silva, Marcelo M. Wanderley, and Gary P. Scavone. On the use of flute air jet as a musical control variable. In *International Conference on New Interfaces for Musical Expression (NIME)*, pages 105–108, 2005.

[17] Michel Debost. *The Simple Flute: From A to Z*. Oxford University Press, 2002.

[18] Diogo Steinke Deconto, Erick Luiz Fontoura Valenga, and Carlos N. Silla Jr. Automatic music score difficulty classification. In *30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 249–256, 2023.

[19] Youssef Ghatas, Magda Fayek, and Mayada Hadhoud. A hybrid deep learning approach for musical difficulty estimation of piano symbolic music. *Alexandria Engineering Journal*, 61:10183–10196, 2022.

[20] Matan Gover and Oded Zewi. Music translation: Generating piano arrangements in different playing levels. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 36–43, 2022.

[21] Ryan Grist. Breathing solutions for woodwind practitioners, 2024.

[22] Susan Hallam and Alfredo Bautista. Processes of instrumental learning: The development of musical expertise. In *The Oxford Handbook of Music Education*, volume 1, pages 657–676. Oxford University Press, 2012.

[23] Yoonchang Han and Kyogu Lee. Hierarchical approach to detect common mistakes of beginner flute players. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 77–82, 2014.

[24] Zakaria Hassein-Bey, Yohann Abbou, Alexandre D'Hooge, Mathieu Giraud, Gilles Guillemain, and Aurélien Jeanneau. What song now? personalized rhythm guitar learning in western popular music. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2025.

[25] Simon Hunt. *Learning to Play the Flute, Volume 1*. Novello, 1979.

[26] Simon Hunt. *Learning to Play the Flute, Volume 2*. Novello, 1981.

[27] Klaus Krippendorff. Computing krippendorff's alpha-reliability. `http://repository.upenn.edu/asc_papers/242`, 2011.

[28] Simon Libřický and Jan Hajič jr. Modeling the difficulty of saxophone music. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2025.

[29] Zofia Mazur and Mariola Łaguna. Assessment of instrumental music performance: Definitions, criteria, measurement. *Educational Research Review*, 115:115–128, 2017.

[30] MuseScore. Musescore pdf importer. `https://musescore.com/import`, 2025. Accessed: 2026-03-10.

[31] Manuel Müllerschön, Anssi Klapuri, Marcelo Rodríguez, and Christian Cardin. Playability prediction in digital guitar learning using interpretable student and song representations. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2025.

[32] Tomoyasu Nakano and Masataka Goto. Using item response theory to aggregate music annotation results of multiple annotators. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[33] Royal Conservatory of Music (RCM). Flute syllabus 2010 edition, 2010.

[34] Wilhelm Popp. *Erster Flöten-Unterricht, Op. 387*. Peters, 1887. Plate C. 37314.

[35] Wilhelm Popp. *Erster Flöten-Unterricht: Op. 387. Flute-Method for Beginners*. Edition Peters; Aug. Cranz, 1887.

[36] Johann Joachim Quantz. *On Playing the Flute*. The Free Press, 1966. First German edition: 1752. Translated by Edward R. Reilly.

[37] Pedro Ramoneda, Vsevolod Eremenko, Alexandre D'Hooge, Emilia Parada-Cabaleiro, and Xavier Serra. Towards explainable and interpretable musical difficulty estimation: A parameter-efficient approach. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[38] Pedro Ramoneda, Dasaem Jeong, Vsevolod Eremenko, Nazif Can Tamer, Marius Miron, and Xavier Serra. Combining piano performance dimensions for score difficulty classification. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 708–715, 2023.

[39] Pedro Ramoneda, Minhee Lee, Dasaem Jeong, J. J. Valero-Mas, and Xavier Serra. Can audio reveal music performance difficulty? insights from the piano syllabus dataset. *arXiv preprint arXiv:2403.03947*, 2024. To appear.

[40] Pedro Ramoneda, Emilia Parada-Cabaleiro, Dasaem Jeong, and Xavier Serra. Difficulty-controlled simplification of piano scores with synthetic data for inclusive music education. *arXiv preprint arXiv:2511.16228*, 2025.

[41] Pedro Ramoneda, Nazif Can Tamer, Vsevolod Eremenko, Xavier Serra, and Marius Miron. Score difficulty analysis for piano performance education based on fingering. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 201–205, 2022.

[42] Pedro Ramoneda, Nazif Can Tamer, Vsevolod Eremenko, Xavier Serra, and Marius Miron. Predicting performance difficulty from piano sheet music images. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 708–715, 2023.

[43] John Sands. *The Complete Flute Player: Omnibus Edition*. Wise Publications, 1996.

[44] Véronique Sébastien, Henri Ralambondrainy, Olivier Sébastien, and Noël Conruyt. Score analyzer: Automatically determining scores difficulty level for instrumental e-Learning. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 571–576, 2012.

[45] Prateek Tandon and Ankit Tandon. Personalized difficulty level classification and feature analysis for guitar tablature. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 747–753, 2015.

[46] Nancy Toff. *The Flute Book: A Complete Guide for Students and Performers*. Oxford University Press, 1996.

[47] Caroline Van Niekerk. Music Education Policy and Implementation: International Perspectives. *Muziki*, 7(2):251–254, November 2010.

[48] Marcel A. Vélez Vásquez, Mariëlle Baelemans, Jonathan Driedger, Willem Zuidema, and John Ashley Burgoyne. Quantifying the ease of playing song chords on the guitar. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 725–732, 2023.

[49] Silvia Winkler, Anne Lohs, Zahavah M. Zinn-Kirchner, Moonef Alotaibi, and Philipp P. Caffier. Tribute to the flute: A literature review of playing-related problems in flautists. *Journal of Multidisciplinary Healthcare*, 17:649–671, 2024.

[50] Haiyun Zhan. Comparing music teaching methods in different countries. In *SHS Web of Conferences*, volume 213, page 02010, 2025.