Interpretable Rule Learning and Evaluation of Early Twentieth-century Music Styles

Christofer Julio[1], Feng-Hsu Lee[2], and Li Su[3]

[1] Social Networks and Human-Centered Computing, Taiwan International Graduate Program
[2] Faculty of Creative Arts, University of Malaya
fenghsulee@um.edu.my
[3] Institute of Information Science, Academia Sinica
lisu@iis.sinica.edu.tw

Abstract. The paper discusses the classification of four music styles, Serialism, Impressionism, Neoclassicism, and Nationalism, of early-twentieth-century music using interpretable rule learning techniques. Three interpretable rule learning techniques are considered: decision tree, minimum description length (MDL) rule list, and rule set (the skope-rule algorithm). The features of the classifiers are fundamental musical elements based on pitch and interval distributions. Objective evaluation based on the F1 score and subjective evaluation using user study is conducted to understand the result of our classifiers from the musicians' point of view. The results show that a rule set is preferred as the algorithm attained the highest scores for objective and subjective evaluations. The rule set can also generate rules which support music theory and provide new insights regarding the musical characteristics of early twentieth-century music.

Keywords: Early twentieth-century music, interpretable AI, rule learning, evaluation, music information retrieval

1 Introduction

The studies regarding the classification task of classical music have undergone major development in the last few years. Multiple machine learning classifiers, from transparent models such as decision trees [12] to black-box models such as support vector machines [12, 11, 24] and more sophisticated neural networks [18, 23, 15, 16, 27], have been utilized and have effectively classified classical music across periods. In addition to focusing on good performance, another research endeavor has focused on interpretability, that is, the extent to which the process by which a model arrives at its decision is transparent and understandable by humans [12, 28].

Despite abundant research on classical music classification, few studies include composers from the early twentieth century. Instead of labeling the early twentieth-century compositions based on their respective styles, most researchers label the composers around this period as "modern" [23, 25]. This "modern" label may not be enough

to become an accurate representation. The early twentieth-century period in classical music consists of various musical genres – each highly distinctive from the others. Musicologists often categorize these styles as *-isms* [4, 7].

Indeed, early twentieth-century music has a few common concepts. For instance, the composers avoid constructed melodies from the previous periods and do not follow the standard tonal harmony [4, 5]. However, the approaches that each style made differ significantly from one another. Serialism, for instance, focuses on utilizing pitch set series or tone rows [4, 1]. Impressionism does not neglect tonality but maximizes the utilization of timbres, layers, underdeveloped motifs, unresolved harmony, and exotic music scales [4, 5]. Nonetheless, Neoclassicism combines the characteristics of music in the previous periods with modern melody and dissonance treatments.[17, 8]. Conducting the classification tasks over these music styles would be insightful due to the unique characteristics of early twentieth-century music and the scarcity of study for this period.

Instead of focusing on performance alone, this study aims more into the interpretability of the result [21]. We want to see whether there is any new insight regarding the characteristics of early twentieth-century music, which may not be found using conventional music analysis. Our research objectives are motivated by the previous studies that have shown the potential to find new insights into classical music, such as the difference in pitch distribution between Mozart and Haydn's string quartet works [11] or the differences in interval utilization between Beethoven and the composers before him, such as Haydn and Bach [12]. On the contrary, the interpretable deep learning approaches for music classification and analysis [13, 26] mostly focus on *post hoc* interpretation [21] over the learned representations and still require decision trees, rules, and linear models to explain it under specific situations [10].Therefore, we take the approach of rule-based, transparent, and *simulatable* (i.e., humans can reason about the entire decision-making process of the model [21]) models instead of black-box ones. Moreover, in this paper, we extensively study various categories of rule-based models, including rule tree, rule list, and rule set [20]. Besides the well-known decision tree, it should be noted that the rule list and rule set models we employed have yet to be considered in music classification problems [22, 9]. For evaluating the result, we perform not only objective evaluation but also subjective evaluation to understand the human's perceptions regarding the rules generated by the models.

To our knowledge, this paper is the first attempt to machine learning classification of early twentieth-century music. This paper has three major aims. First, we propose a new dataset regarding early-twentieth-century music in symbolic format. Second, we investigate various rule-based machine learning models for music style classification on this new dataset. Lastly, the interpretability of the classification results and the selected rules and features are analyzed and discussed with both objective and subjective aspects.

2 Data

Since the repertoires of early twentieth-century music are wide and complicated, we imposed several restrictions in choosing the works for the dataset. The subject of this research is limited to early twentieth-century composers' piano works, as we tend to

Table 1: The proposed dataset for classification of the early 20th-century music styles. The number of samples of each composer and each style in the dataset are shown.

Styles	Composers	# of samples		Styles	Composers	# of samples	
Serialism	Arnold Schoenberg	22	48	Neoclassicism	Maurice Ravel	11	104
	Alban Berg	2			Paul Hindemith	86	
	Anton Webern	5			Béla Bartók	7	
	Hanns Eisler	19		Nationalism	Béla Bartók	247	304
Impressionism	Claude Debussy	87	110		Leoš Janáček	43	
	Maurice Ravel	23			Manuel de Falla	14	

use homogeneous data to avoid any potential problems related to instrumentation. This approach has also been used in previous studies, where the researchers limited their choice of instruments to only string quartet [11, 14, 24], the melody of the violin [6], piano solo [23, 25], and orchestra [25]. We included the first 20 measures (approximately one page) of every composition to prevent the imbalance of the dataset. Besides, we chose the styles and composers based on the number of piano pieces for each composer and the availability of the scores. Composers who only have a few piano works were not selected. In addition, we only choose the works in the public domain. Hence, the dataset consists of four styles and ten composers, see Table 1.

Except for Béla Bartók and Maurice Ravel, each composer's compositions are classified in one style. Ravel's works are divided into Impressionism and Neoclassicism, as Ravel had distinctive styles during his early and late period [4]. In addition, Bartók's works with Nationalism style are chosen manually based on existing literature due to his unique approaches between the traditional and modernistic style [4]. Besides Nationalism, a few of Bartók's piano works are also separated into Neoclassicism due to the use of classical forms. Other Bartók's piano works, which do not fall into these two styles (such as Night music), are not included. Lastly, the pre-serialism works from Serialism composers, such as Schoenberg's late romantic works, are not incorporated into the dataset.

The dataset of the early twentieth-century music in this study utilizes note events derived from musicXML, as it can save more information compared to MIDI. We collected the data from the Petrucci music library (imslp.org) and manually converted them to the MusicXML format. Note events, including pitch value, onset time, and duration, are then extracted. The dataset will be publicly announced after this paper is accepted.

3 Method

3.1 Data representation

We consider pitch-related features and intervals as the data representation, as our pilot study demonstrated that they are more relevant than other music features for classifying our early twentieth-century dataset. [2]. Given a music piece $\{x_i\}_{i=1}^{N}$ with N notes, the pitch value (in MIDI number) of the ith note being p_i, the pitch range (r_p), pitch mean

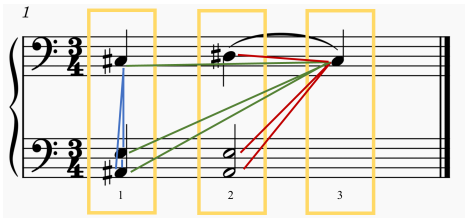

Fig. 1: The example of horizontal and vertical interval feature calculation. Excerpt taken from Bartók's *9 Little Pieces for Piano no.5*.

(μ_p), and pitch standard deviation (σ_p) over all time steps are

$$r_p := \max_i(p_i) - \min_i(p_i)\,; \quad \mu_p := \frac{1}{N}\sum_{i=0}^{N} p_i\,; \quad \sigma_p := \sqrt{\frac{\sum (p_i - \mu_p)^2}{N}}\,. \tag{1}$$

Then, vertical interval features are calculated to understand the harmony of the repertoire. For simplicity, the positions of notes are grouped based on the beats. By normalizing all the data to 4/4 meter, a *beat* refers to all notes within a quarter note duration. For example, given a music excerpt with Y beats, the position of a note x_i is y, $1 \leq y \leq Y$, if its onset is in the interval $[y, y+1)$, i.e., between the yth and the $(y+1)$th beat of the music piece. For any two notes x_i and x_j at the same beat y, assuming $p_i \geq p_j$, the vertical interval between x_i and x_j at y is 12 if $p_i - p_j = 12n$, $n \in \mathbb{N}$, and is $p_i - p_j \pmod{12}$ for other cases. That means a vertical interval is a value ranging from 0 (unison) to 12 (perfect octave). The distribution of the vertical intervals over all time steps is then represented as a 13-dimensional vector, obtained by aggregating the counts of each interval class over all the time steps. The final vertical interval feature (denoted as $\bar{v}$) is a min-max scale normalization over this distribution.

In addition, we employ another feature based on the horizontal interval for understanding the relationship between neighboring notes. For the horizontal features, we consider two groups of notes by m beats apart from each other. where Following the *skip-gram* technique in the field of natural language processing, m is the number of *skips*, and $m = 0$ represents no skip. Similar to vertical interval, the horizontal interval of x_i and x_{i+m+1} (assuming $p_i > p_{i+m+1}$), is 12 if $p_i - p_{i+m+1} = 12n$, $n \in \mathbb{N}$, and is $p_i - p_{i+m+1} \pmod{12}$ for other cases. Similar to the normalized distribution of vertical intervals, the normalized distribution of m-skip horizontal intervals (denoted as $\bar{h}^{(m)}$) is also a 13-dimensional vector by aggregating the counts of each interval and min-max normalization.

The straightforward way of calculating the vertical and horizontal interval features is demonstrated by an example in Figure 1. There are three beats (indicated by the yellow boxes) in this example. The vertical interval calculations are shown in the blue lines. At the first beat, for the intervals from the note C♯3, we calculate every possible interval in the same time stamp. Here, calculations are made from C♯3 to E3 (i.e., a minor third, also denoted as "V-m3") and from A♯2 to C♯3 (i.e., minor third or V-m3). The rest of the notes are treated similarly without repetitions; for example, at this beat,

we also have an interval between A♯2 and E3 (diminished fifth or V-d5). Summing up the vertical intervals over all the timestamps in Figure 1, we have in total one V-m2, two V-m3, two V-d5 and one V-P5, so the distribution is $[0, 1, 0, 2, 0, 0, 2, 1, 0, 0, 0, 0, 0]$ and the min-max-normalized distribution is $\bar{v} = [0, 0.5, 0, 1, 0, 0, 1, 0.5, 0, 0, 0, 0, 0]$. Meanwhile, the red lines show the calculation of horizontal intervals without any skip. Our example here calculates the horizontal intervals between the second and the third beats, which result in three intervals: D♯3 to C♯3 (major second, or denoted as H-M2), C♯3 to E3 (H-m3), and A♯2 to C♯3 (H-P5). Lastly, the green line indicates the horizontal interval calculation with skips. In this example, we only demonstrate the interval calculation with one skip, i.e., between the first and third beats. The calculation results in three intervals: C♯3 to C♯3 (H-P1), C♯3 to E3 (H-m2), and A♯2 to C♯3 (H-m3). The method of summing up different timestamps is similar to the case of vertical intervals.

In the remainder of this paper, the number of skips is not specified if it is zero. To summarize, we consider the pitch features (pitch range, pitch mean, and pitch standard deviation, totaling three dimensions), vertical interval features (13 dimensions), and horizontal interval features with skips from 0 to 2 ($13 \times 3 = 39$ dimensions). This results in a total feature dimension of 55.

3.2 Classifiers

We consider three categories of rule-based algorithms: rule trees (i.e., decision trees), rule lists, and rule sets. These three are interpretable in that they are all constructed with conditional statements (i.e., if-then-else rules) of the input features and the corresponding outcomes [10]. In decision tree, the if-then-else rules form a tree structure in which the internal nodes represent conditions of features, and each leaf node represents a class label. In rule lists and rule sets, each condition of an if-then clause can incorporate multiple input variables. Specifically, in rule lists, rules are the conditions ordered in nested if-else statements, while in rule sets, rules are unordered and independent from each other in that the else statements do not connect the rules [10]. As for visual representation, rule trees are often illustrated in tree graphs, while rule lists and rule sets tend to have textual or tabular representation. Hence, rule trees, rule lists and rule sets are not equivalent and are different in multiple aspects.

The rule tree classifier we adopt is the decision tree with the optimized CART Algorithm [3], available from the scikit-learn library.[4] For the rule list classifier, we utilize the minimum description length (MDL) rule list, a probabilistic multi-class classifier algorithm. MDL rule list is designed using the minimum description length principle, which chooses the best model based on the ability to compress the data [22].[5] The MDL Rule list requires only a few hyperparameters to work and can acquire competitive accuracy [22]. Lastly, for the rule set classifier, we utilize skope-rules.[6] Similar to Rulefit [20], the rules from Skope-rules are chosen by extracting the path of the tree from multiple decision trees. However, the difference lies in establishing the final rules.

[4] https://scikit-learn.org/stable/modules/tree.html
[5] https://github.com/HMProenca/MDLRuleLists
[6] Source code available at https://github.com/scikit-learn-contrib/skope-rules

Table 2: Classification results using the three rule-based classifiers on the four styles of early 20th-century music. Precision (P), recall (R) and F1-score (F1) values are shown.

	rule tree			rule list			rule set		
	P	R	F1	P	R	F1	P	R	F1
Serialism	0.80	0.59	0.68	0.68	0.68	0.68	0.99	0.73	0.84
Neoclassicism	0.60	0.60	0.60	0.52	0.42	0.47	0.84	0.56	0.67
Impressionism	0.56	0.57	0.57	0.55	0.27	0.37	0.68	0.63	0.66
Nationalism	0.62	0.67	0.64	0.57	0.97	0.71	0.83	0.58	0.68
Average	0.65	0.61	0.62	0.58	0.59	0.56	0.84	0.63	0.71

Table 3: The extracted rule set of four classes

Rules 1	Rules 2	Rules 3	Rules 4	Class
Pitch Range $>$ 45.5	H-M7 (2 skip) $>$ 0.17	V-m7 $\leq$ 0.39	V-P8 $\leq$ 0.8	Impressionism
Pitch Range $\leq$ 52.5	V-P1 $>$ 0.15	V-P8 $>$ 0.23	H-m2 $>$ 0.05	Nationalism
H-P8 $\leq$ 0.004	H-M7 (1 skip) $>$ 0.05	H-M6 (2 skip) $>$ 0.15	VI 4 $>$ 0.49	Serialism
H-A4 $\leq$ 0.5	H-P8 $>$ 0.01	V-m7 $>$ 0.4	V-M7 $>$ 0.3	Neoclassicism

Skope-rules filter the rules using out-of-bag (OOB) precision and recall thresholds and the semantic deduplication method for maintaining the diversity of the rules [20].

The hyperparameters utilized in this study are described as follows. For decision tree, we use a maximum depth of four, gini impurity as the criterion, and minimum sample split as two. The rest are followed by the default settings of Scikit-learn's Decision Tree. Meanwhile, the parameters used for the MDL rule list classifier is elaborated as follows: static data discretization, the maximum size of each rule description being 4, the number of cut point of each variable being 1, minimum support being 0.1, and alpha gain being 0. Lastly, for Skope-rules, we utilize similar hyperparameters with Decision Tree, except that we limit the estimated number of the generated tree to 72.

4 Experiments

4.1 Experimental settings

Before training the classifiers, data augmentation has to be done since the size of the dataset is considered small. In this case, a normalization process is performed as suggested by [11], converting every sample's key into C major and a minor. However, these adjustments are strictly for tonal music, and this conversion step is skipped for atonal works. Then, we perform pitch shifting from -5 to 6 semitones for each work. To balance the dataset, we randomly select 35 percent of Nationalism samples due to their larger number. The dataset is then divided into training and testing sets with the 80:20 ratio. Lastly, considering the number of samples, 5-fold cross-validation (CV) is performed for each experiment to obtain stable classification results. The test-set precision, recall, and F1-score values averaged over the 5-fold CV are reported and compared.

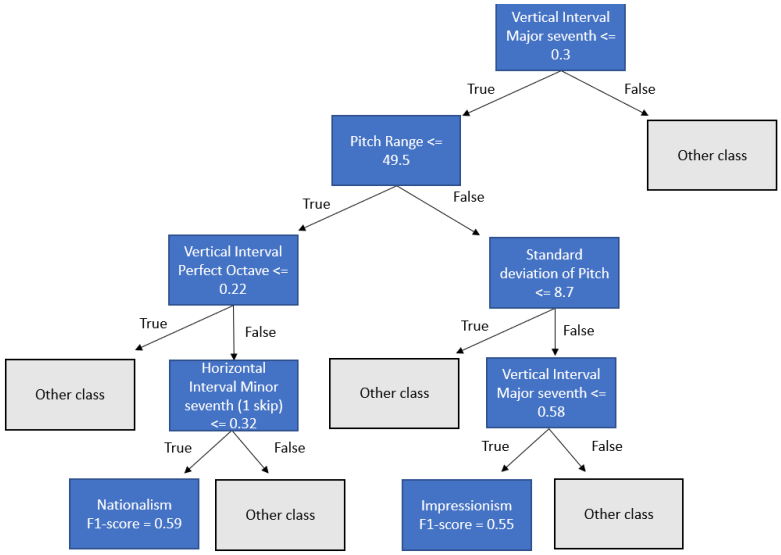

Fig. 2: Partially extracted tree of Nationalism and Impressionism class. ”Other class” denotes the weak or irrelevant class which is neither Nationalism nor Impressionism.

↘19 list of ELSE IF with low impact results

➢ELSE IF Pitch Range < 43.0 AND Vertical Interval Perfect Octave >= 0.36 AND Horizontal Interval Perfect Fourth (2 skip) >= 0.4 AND Vertical Interval Minor Second < 1.0 THEN Probability of Nationalism = 1.0

↘12 list of ELSE IF with low impact results

➢ELSE THEN Probability of Impressionism = 0.96; Probability of Neoclassicism = 0.04

Fig. 3: The extracted rule list of Nationalism and Impressionism Class. The hidden lists contain other n rules which have low impacts to the classification decisions.

4.2 Objective evaluation

Table 2 shows the classification result of decision tree, MDL rule list, and Skope-rules. Skope-rules achieves an average F1-score at 0.71, outperforming both decision tree (F1-score = 0.62) and MDL rule list (F1-score = 0.56) by a wide margin. Meanwhile, we have slightly different results for the F1-score of each class. Skope-rules still dominate in Serialism, Neoclassicism, and Impressionism classes, followed by decision tree. However, for Nationalism, the MDL rule list achieves a better result than the other two, with F1-score = 0.71. Lastly, the results of both Neoclassicism and Impressionism of the MDL rule list are underwhelming, with the F1-score less than 0.5.

4.3 Subjective evaluation

A user study in the form of a questionnaire is utilized to understand the interpretability of the models for musicologists, musicians, composers, and other music-related

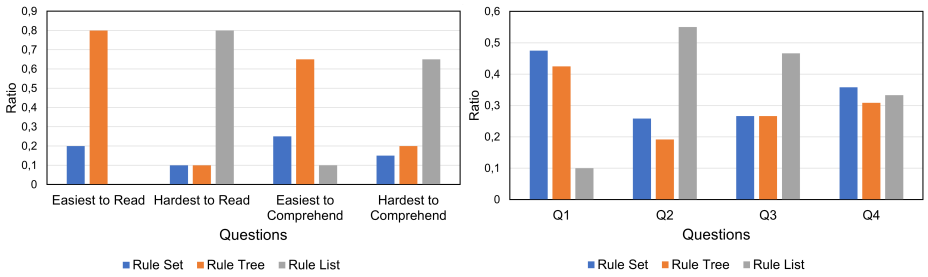

Fig. 4: Results of the subjective tests. Left: visualization test. Right: content test.

researchers. The questionnaire of our subjective evaluation contains two parts, *visualization test* and *content test*. The first part evaluates the subjective response to the *visualization* quality of the rules. Rule tree is visualized with a decision tree graph (e.g., Figure 2) while rule list is represented by text, consisting of the if, else-if, and else rules (e.g., Figure 3). Lastly, rule set is represented by a table (e.g., Figure 3). The questions then aim to understand the most favorable representation of the visualization results based on the opinions of respondents. Since the 5-fold CV generates five different lists, trees, and sets, we decided to take the best graphs or rules based on the best F1-score of the folds for the questionnaire questions.

The second part of the subjective test evaluates the *content* of the generated rules of each class. In the second part of the subjective test, we present $C_2^4 = 6$ question sets of rule tree, rule list, and rule set based on the binary classification; the six sets contain each of the two styles selected from the four musical styles for pairwise comparison. Each set consists of four questions. They are (Q1) From the three options, which one gives the best result according to current music theory? (Q2) From the three options, which one gives the worst result according to current music theory? (Q3) From the three options, which one gives the most unusual rules? (Q4) From the three options, which one gives the least unusual rules?

20 participants joined the subjective test. 18 of them have a degree in music. Among the participants, 13 have more than 11 years of experience in music. On a scale of 1-5, 7 participants are very familiar with early twentieth-century music (scale 4-5), while 10 participants are familiar with early twentieth-century music (scale 3). Only 3 participants are quite unfamiliar with early twentieth-century music (scale 2).

The left-hand side of Figure 4 shows the result of the subjective test. For the visualization test, rule tree is the model representation that is easiest to read, followed by rule set. Meanwhile, rule list is the hardest to read. Similarly, among the three models, the rule tree is also the most comprehensible, followed by rule set and rule list.

The right-hand side of Figure 4 shows the result of the content. In line with the result of the visualization test, the answers to Q1 and Q2 of the content test show that most participants favor rule set and rule tree over rule list. However, unique results are seen based on the answer of Q3. Even though rule set has the highest ratio in Q1, it turns out rule set also occupies second place in Q3. It means that although rule set has rules strongly similar to current music theory, some are also considered unusual.

5 Discussion

5.1 The Subjective and Objective Evaluation Analysis

The objective and subjective evaluations conducted in this study show several similar trends. The visualization and content test show identical results regarding the most accurate classifier among the three representations. Skope-rules appears to be the best classifier with the F1-score = 0.71, and the classifier shows the best result for the current music theory based on the *content* test (see Q1 and Q2 in Figure 4). Meanwhile, the rule tree comes second with F1-score = 0.62, with the second-best accuracy towards the current music theory. On the other hand, the rule list becomes the worst classifier among these trees with the lowest F1-score, lowest Q1, and highest Q2 value of *content* test. The Q3 answers show that rule list generates the most unusual rules compared to others. There may be two possible explanations regarding this matter. First, the unusual rules may be the signs of the new possible finding regarding the theory of musicology. Second, the rules from rule list may not be accurate because it occupies the lowest F1-score in objective evaluation. However, at the current state of the study, we are unable to identify whether these found rules from the rule list are truly insightful, and further investigations are required. Lastly, based on the Q4 of the content test, no clear trend was found.

Meanwhile, regarding the interpretability of the rules, we still observe contrasting outcomes in between visualization tests Based on the result of the visualization test (Figure 4), rule tree offers better comprehensibility and readability compared to rule list and rule set. Rule set comes second despite having the highest precision, recall, and F1-score on the objective evaluation. The result in our case shows that a higher F1-score does not always imply better interpretability. This is possibly due to data representation: the tree data structure in rule tree has the advantage of showing the relationship between the classes. For instance, in Figure 2, the readers can easily notice the distinctions of Nationalism and Impressionism classes directly from the ramification on the first depth onward. Meanwhile, for the rule list and rule set, the readers need to compare each rule one by one. In addition, the rules generated from the rule tree always show at least one related feature of both classes (see Figure 2) since, in the tree model, two child nodes always have at least one shared parent node. On the contrary, in both rule list and rule set, there are possibilities that all features of both classes are distinctive. Readers may be confused in comparing the rules if all the rules between classes are unrelated.

The rule list shows the most inferior performance from both the subjective and objective perspectives: The average F1-score is only 0.56 (although it performs the best in Nationalism), and it is the hardest to read, comprehend, and the worst according to music theory. This might be due to data representation: there is a possibility that even though the rule list may produce reasonable rules, the subjective evaluation participants tend to choose other models due to the unfamiliarity of the respondents with the IF-ELSE concept in Figure 3, which are computer science rather than musical knowledge.

Based on the subjective and objective evaluation results, rule set shows the best accuracy in the F1-score and the content test while the outcome of the visualization test still indicates the potential of the rule tree as a good representation that favors music practitioners and musicologists. Besides, the results of the rule list are least favorable.

Table 4: The features of the four random-chosen excerpts. The green color shows that the feature fits the rule set and the red color shows that the feature unfits the rule set.

	Example (Composer)	C. Debussy	B. Bartók	A. Schoenberg	P. Hindemith
	Class	Impressionism	Nationalism	Serialism	Neoclassicism
Impression	Pitch range $>$ 45.5	73	43	63	72
	H-M7 (2 skip) $>$ 0.17	0	0.12	0.13	0
	V-m7 $\leq$ 0.39	0.22	0.06	1	0.5
	V-P8 $\leq$ 0.8	1	0.16	0	0.74
National	Pitch range $\leq$ 52.5	73	43	63	72
	V-P1 $>$ 0.15	0.007	0	0	0
	V-P8 $>$ 0.23	1	0.16	0	0.74
	H-m2 $>$ 0.05	0.17	0.62	0.99	0.39
Serialism	H-P8 $\leq$ 0.004	0.74	0	0	0.13
	H-M7 (1 skip) $>$ 0.05	0	0.02	0.3	0
	H-M6 (2 skip) $>$ 0.15	0.46	0.09	0.29	0.12
	V-M3 $>$ 0.49	0.65	0.35	0.53	0.84
Neoclassic	H-A4 $\leq$ 0.5	0.21	0.38	0.97	0.11
	H-P8 $>$ 0.01	0.74	0	0	0.13
	V-m7 $>$ 0.4	0.22	0.06	1	0.5
	V-M7 $>$ 0.3	0.03	0.19	0.54	0.39

5.2 Case Study

In this part, we perform a case study to see how the learned rules work on real-world music examples. We randomly select four excerpts from our dataset to represent each respective style. The music pieces are the excerpts chosen from Claude Debussy's *Nocturne*, Béla Bartók's *Nine Little Pieces for Piano*, Arnold Schoenberg's *Suite for Piano* and Paul Hindemith's *Ludus Tonalis*.

The rule set on Table 3 is utilized in the case study since rule set is the most recommended algorithm according to our previous discussion. The details of the chosen excerpts and the values of those features which appear in the rules on Table 3 can be seen in Table 4. Although the statements of feature in the rule set are combined with the logical AND, we discuss the feature separately for convenience. For example, the pitch range of Debussy's *Nocturne* fits the rule "Pitch range $>$ 45.5" for Impressionism since its value is 73. In addition, the pitch range of Bartók's piece also fits this rule since it is non-Impressionism, and its value, 43, is smaller than 45.5. As a result, the value with the green color in Table 4 indicates that it fits the rule description of the music style, while the value with the red color shows that it does not fit the rules.

The results in Table 4 indicate promising outcomes. For Debussy's *Nocturne* (Impressionism) and Bartók's *Moderato* (Nationalism), two out of the four rules predict the style label correctly Meanwhile, for Schoenberg's *Präludium* (Serialism) and Paul Hindemith's *Preludio* (Neoclassicism), all the rules are correct. It should be noted that current music theory does support certain generated rules on Table 3. For instance, the pitch range of Impressionism is supposed to be larger than 45.5 semitones, and Debussy's *Nocturne* fulfills the requirements. The requirement of such a wide pitch range

may be correlated with the main characteristics of Impressionism piano works; among them are "open chord, wide spacing, and extreme register" [19, p. 169].

Meanwhile, some other results demonstrate unusual rules based on the perspective of musicology. For instance, the rule from the rule set shows the importance of V-M3 (major third) in Serialism composition. However, Serialism works do not stress the utilization of the major third since Serialism composition utilizes intervals based on the tone rows [4]. Lastly, another issue that concerns us is that some rules generated by the rule set are very weak. For instance, one of the Neoclassicism rules states that the normalized value of H-P8 (perfect octave) needs to be larger than 0.01, a very small lower threshold. Therefore, such a rule may not be as insightful since any music piece that merely utilizes a few numbers of the horizontal perfect octave interval might satisfy it. Hence, in certain parts, such rules do not highlight the important characteristics of the styles but are redundant. However, certain weak rules are still able to show some insights. For example, the rule H-M7 (skip 1) > 0.05 seems weak given that 0.05 is a small lower threshold. However, based on observations, the excerpts in the other three styles have this feature smaller than 0.05, meaning that the horizontal major seventh interval with skip 1 rarely appears except in Serialism.

6 Conclusion

To conclude, the systematic study of rule-based interpretable algorithms for classifying the styles of early twentieth-century music indicates that the rule-set-based algorithm, Skope-rules, shows the best performance in the precision, recall, and F1-score of the objective evaluation and also offers decent comprehensibility and consistency of music theory in the subjective tests. In addition, the chosen algorithm with our feature design can find the rules in line with the current music theory, as well as the promising unusual rules which may show new insights regarding early twentieth-century music. Rule tree, on the other hand, is able to provide the best result in visualization test, yet is unable to outperform rule set in other evaluation sections. Thus, while the previous studies on rule-based interpretable music AI mostly considered decision trees, we suggest using rule-set-based algorithms for related research directions.

References

1. Auner, J.: A Schoenberg Reader: Documents of a Life (2003)
2. Author, A.: (Anonymized for double-blind review). Master's thesis, Anonymous school (2022)
3. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: CART: Classification and regression trees. Routledge (1984)
4. Burkholder, J.P., Grout, D.J., Palisca, C.V.: A History of Western Music: Tenth Edition. New York, NY: W.W Norton & Company. (2019)
5. Clark, N.A., Heflin, T., Kluball, J., Kramer, E.: Understanding music: Past and present. University System of Georgia, University Press of North Georgia (2015)
6. De Carvalho Junior, A.D., Batista, L.V.: Composer classification in symbolic data using ppm. In: 11th International Conference on Machine Learning and Applications. vol. 2, pp. 345–350. IEEE (2012)

7. Frisch, W.: Music in the Twentieth and Twenty-First Centuries. New York, NY: W.W Norton & Company. (2013)
8. Gabriela, V.: Baroque reflections in Ludus Tonalis by Paul Hindemith. In: 11th WSEAS International Conference on Acoustics & Music: Theory & Applications (AMTA'10). vol. 1 (2010)
9. Gardin, F., Gautier, R., Goix, N., Ndiaye, B., Schertzer, J.M.: Skope-rules. https://github.com/scikit-learn-contrib/skope-rules
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5) (aug 2018). https://doi.org/10.1145/3236009, https://doi.org/10.1145/3236009
11. Herlands, W., Der, R., Greenberg, Y., Levin, S.: A machine learning approach to musically meaningful homogeneous style classification. In: AAAI Conference on Artificial Intelligence. vol. 28 (2014)
12. Herremans, D., Martens, D., Sörensen, K.: Composer classification models for music-theory building. In: Computational Music Analysis, pp. 369–392. Springer (2016)
13. Kelz, R., Widmer, G.: Towards interpretable polyphonic transcription with invertible neural networks. arXiv preprint arXiv:1909.01622 (2019)
14. Kempfert, K.C., Wong, S.W.: Where does Haydn end and Mozart begin? Composer classification of string quartets. Journal of New Music Research **49**(5), 457–476 (2020)
15. Kim, S., Lee, H., Park, S., Lee, J., Choi, K.: Deep composer classification using symbolic representation. arXiv preprint arXiv:2010.00823 (2020)
16. Kong, Q., Choi, K., Wang, Y.: Large-scale MIDI-based composer classification. arXiv preprint arXiv:2010.14805 (2020)
17. Kostka, S.: Materials and techniques of twentieth-century music, 3rd ed. Pearson Prentice Hall (2006)
18. Micchi, G.: A neural network for composer classification. In: International Society for Music Information Retrieval Conference (ISMIR) (2018)
19. Miller, H.M.: History of Music. Barnes & Noble Books (1973)
20. Molnar, C.: Interpretable machine learning: A Guide for Making Black Box Models Explainable. Lulu. com (2019)
21. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences **116**(44), 22071–22080 (2019)
22. Proença, H.M., van Leeuwen, M.: Interpretable multiclass classification by mdl-based rule lists. Information Sciences **512**, 1372–1393 (2020)
23. Van Spijker, B.: Classifying classical piano music into time period using machine learning. Master's thesis, University of Twente (2020)
24. Velarde, G., Weyde, T., Chacón, C.C., Meredith, D., Grachten, M.: Composer recognition based on 2d-filtered piano-rolls. In: International Society for Music Information Retrieval Conference (ISMIR). pp. 115–121. International Society for Music Information Retrieval (2016)
25. Weiß, C., Mauch, M., Dixon, S., Müller, M.: Investigating style evolution of Western classical music: A computational approach. Musicae Scientiae **23**(4), 486–507 (2019)
26. Won, M., Chun, S., Serra, X.: Toward interpretable music tagging with self-attention. arXiv preprint arXiv:1906.04972 (2019)
27. Yang, D., Tsai, T.: Composer classification with cross-modal transfer learning and musically-informed augmentation. In: International Society for Music Information Retrieval Conference (ISMIR). pp. 802–809 (2021)
28. Yu, H., Varshney, L.R., Garnett, G.E., Kumar, R.: Learning interpretable musical compositional rules and traces. In: ICML Workshop on Human Interpretability in Machine Learning (WHI) (2016)