MODELING RAGTIME STRUCTURE AND STYLE FROM RHYTHMIC NOTATION ALONE

Phillip B. Kirlin
Department of Computer Science, Rhodes College
kirlinp@rhodes.edu

ABSTRACT

We explore the extent to which rhythmic features alone, specifically binary note-onset patterns and the Longuet-Higgins and Lee (LHL) syncopation metric, can predict both musical phrase boundaries and composer identity in ragtime music. Using a novel dataset of 71 hand-labeled symbolic scores by Scott Joplin, James Scott, and Joseph Lamb, we present statistical analyses revealing that rhythmic and syncopation patterns differ significantly across composers and metrical positions within a 16-measure strain. We evaluate how effectively these rhythmic patterns predict ragtime phrase boundaries by training and comparing multiple machine learning models, including logistic regression, support vector machines, and neural network architectures such as convolutional neural networks, bidirectional long short-term memory networks, and transformers. An ablation study confirms that the LHL syncopation measure significantly improves predictive performance. We further demonstrate that these same rhythmic features can be used to classify composer identity either from an entire piece or from a 16-measure strain, and discuss the musicological interpretations of these findings. These results provide clear evidence that the rhythmic aspect of musical notation alone can model structural and stylistic differences in compositions.

1. INTRODUCTION

Understanding musical structure is a central challenge in computational musicology. While prior work has approached this problem from melodic, harmonic, and rhythmic perspectives, most systems aim to leverage whichever features are most predictive, regardless of their musical origin. Yet in certain genres, such as ragtime, rhythmic style is not merely a contributing factor, but a defining element of the genre.

Ragtime is characterized by its distinctive rhythmic features, particularly its syncopated or "ragged" rhythms. During ragtime's heyday in the late 1890s and early 1900s, the stylistic convention emerged of structuring such compositions into 16-measure phrases or *strains*; composers such as Scott Joplin, James Scott, and Joseph Lamb best exemplified this convention [1, 2]. This standardized phrase structure, combined with the characteristic syncopated rhythmic complexity, provides an ideal context for investigating how rhythmic features might encode both structural and composer-specific stylistic information.

In particular, we ask to what extent the structure of ragtime music can be understood and predicted using the rhythmic aspect of notation alone. While a number of recent studies have attempted to quantify the way syncopation — a purely rhythmic phenomenon — is used in ragtime compositions, previous research has not systematically investigated syncopation's potential to predict structural elements such as phrase boundaries or composer identity. Understanding how rhythmic and syncopation patterns encode such information provides potential insights into cognitive processing of rhythm and can inform other areas of computational musicology.

In this paper, we directly investigate the relationship between syncopation and the development of musical phrases and strains in a ragtime composition. We then extend the investigation to study if different ragtime composers utilized syncopation differently within musical phrases. We show that we can use syncopation along with note onset patterns — no other melodic or harmonic information — to reliably predict both phrase boundaries and composer identity in symbolic ragtime scores. This is made possible by a newly-released symbolic dataset of 71 ragtime piano compositions, annotated with composer labels and hand-labeled 16-measure segment boundaries.

Our contributions are as follows. We publicly release a new dataset of ragtime music annotated with 16-measure phrase boundaries and composer labels. We present a detailed analysis of how syncopation, measured using the Longuet-Higgins and Lee (LHL) metric [3], varies within and across pieces, composers, and phrase locations. Our statistical results confirm that composers use syncopation in measurably distinct ways within phrases. We use machine learning models to predict the location of segment boundaries using rhythmic onset patterns and syncopation values alone. We show that these same rhythmic features can reliably predict composer identity, both at the segment level and the full-piece level, and we provide a musicological interpretation of these findings. Lastly, we confirm the impact of the syncopation factor with an ablation study. The dataset and code for all of the experiments described here is publicly available. [1]

[1] https://github.com/pkirlin/ragtime-tenor-2025

2. RAGTIME, SYNCOPATION, AND MUSICAL PHRASES

Syncopation in music occurs when a listener, expecting to hear the musical onset of a note on a strong beat, instead finds that onset shifted to a weak beat, subverting their expectations [4, 5]. While syncopation is found in many musical genres, it is particularly identified with ragtime [2]. Because of this close association, it is natural to study how the use of syncopated patterns varies in the genre. Music historians have noted that the particular varieties of syncopated patterns have varied in usage over time [1, 6, 7], and computational studies have confirmed this [8, 9, 10, 11]. Syncopation can be measured through various metrics, including the Longuet-Higgins and Lee (LHL) metric [3], various schemes utilizing inner metric analysis [12, 13] or the Sioros-Guedes metric [14]. We rely on the LHL metric here because it closely aligns with human perception [15] and has been used in other corpus studies of ragtime [10, 9]. We describe the calculation of this metric in the following section.

A prototypical ragtime composition has a set musical structure consisting of three or four sections of music known as *strains*, each of which is sixteen measures long. Each strain is typically denoted by a letter of the alphabet, "A," "B," etc. Strains are often repeated and reprised. Each strain is usually divided into four phrases of four measures each, occasionally with the third phrase a repetition of the first. While the majority of a composition consists of these strains, frequently there will be other musical phrases included that are not a part of any strain, such as an introduction, short interludes, or a coda.

3. CREATION AND ANALYSIS OF THE DATASET

There are many datasets of ragtime music available [8], but no symbolic ones with strain boundaries clearly identified. Therefore, we set out to create a dataset of "classic rags," a term which refers to a ragtime piano composition written in the style of Scott Joplin (1867 or 1868–1917) that usually follows a prescribed musical form [16]. These compositions are written in duple meter (usually 2/4 time, though some in 4/4 time) and with well-defined 16-measure strains. "Classic rags" specifically exclude ragtime dances, waltzes, or marches. The largest set of classic rags was written by the "big three" ragtime composers of Joplin, James Scott (1885–1938) and Joseph Lamb (1887–1960), who stand out as best exemplifying the ragtime genre [2].

Creating the Dataset. We created this dataset by examining the works of Joplin, Scott, and Lamb, and selecting all the compositions that fit the classic rag format. We then located symbolic scores in either Humdrum or MusicXML format, and manually located each 16-measure musical strain of each composition, assigning them the letters "A," "B," etc., along with also notating variations of each strain. We use the term "variation" here to mean any repetition of the strain that is not identical to the original occurrence. For every musical phrase *not* part of a 16-measure strain, we assigned a label such as "introduction," "interlude," or "coda."

In the end, we collected 71 compositions by the big three composers (31 by Joplin, 28 by Scott, and 12 by Lamb). All were in 2/4 time except for two in 4/4 time (one each by Joplin and Lamb). These compositions consisted of 10,414 total measures of music. There were a total of 604 16-measure strains in the corpus, including repeated strains, totaling 9,664 measures of music. The phrase annotations were stored in a text file recording the name of the composition, the composer, the name of the phrase or strain, the variation number of the phrase or strain, and its length in number of measures. Every measure of every composition therefore falls into exactly one phrase or strain, which we will collectively call *segments*.

Preprocessing. Classic piano rags feature a syncopated melody in the right hand paired with a largely non-syncopated left-hand accompaniment [1]. Because our analysis focuses exclusively on the rhythmic aspect of music notation, we began by converting each measure of the right-hand melody parts of the compositions into a *binary onset pattern*, a standard representation in studies of ragtime syncopation [8, 10, 9]. A binary onset pattern is a pattern of ones and zeros where a one represents a note onset and a zero represents a lack of a note onset at a particular subdivision of the beat. For pieces in 2/4 time, we computed these patterns at the 16th-note level of granularity, meaning a measure of four eighth notes would be represented as "`10101010`." For pieces in 4/4 time, we computed these patterns at the 8th-note level, as it was clear from the notation in these compositions that the underlying quarter note pulse in the 4/4 pieces was equivalent to the eighth note pulse in the 2/4 pieces. This step ensured that every measure in the entire dataset was represented by a pattern of eight ones and zeros.

Using these patterns, we computed the LHL syncopation metric for each measure. This metric is zero for measures with no syncopation, and increases for each occurrence of a note onset on a weak beat followed immediately by the lack of an onset on the following strong beat. Syncopations crossing longer divisions of the measure increase the metric more. For instance, the pattern `01010101` contains three instances of a weak beat "`1`" followed by a strong beat "`0`." The first and last have LHL values of 1, while the middle instance has a value of 2 because it crosses the midpoint of the measure. Therefore, this measure has an overall LHL value of 4. If this measure were followed by a measure with no note onset on the downbeat, the LHL value would increase by 3 (to a total of 7) for the additional syncopation crossing the barline.

As the final step of preprocessing, we calculated the LHL value from each previously-computed binary onset pattern. Our final dataset consisted of every measure of music from the melodies in the corpus, aligned with the corresponding ragtime segment annotations, binary onset pattern, and LHL value.

Analysis. Previous work has shown that the big three ragtime composers used more syncopation than their contemporaries [9] and that their use of syncopation varied depending on interactions with the metrical hierarchy [11].

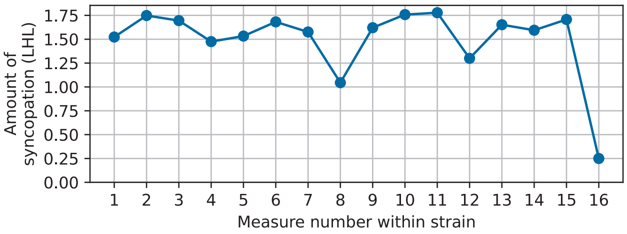

Figure 1. Average syncopation values during each measure of a 16-measure strain.

Here, we examine syncopation specifically within 16-measure strains.

We first analyzed the change in LHL syncopation values during a 16-measure strain. We isolated only these strains from our dataset and averaged the LHL values for each measure across all of the strains. Figure 1 illustrates that these values are not constant. We can see how syncopation tends to fall at natural points of rest within the strain, namely at the end of the strain (measure 16), halfway (measure 8), and one-quarter and three-quarters of the way through (measures 4 and 12). This should not be surprising — these are the concluding points of the natural four-measure phrases within the strain, where one would expect a release of tension. The average LHL values at measures 3–4, 7–8, 11–12, and 15–16 clearly show this release.

We ran a one-sided permutation test to examine if the average LHL value during a strain rises and falls due to deliberate variations in the use of syncopation or due to chance. We measured the total variability in the average LHL values during a strain by calculating the sum of the squared differences between adjacent measures. To generate a null distribution, we randomly permuted the measures within each strain, thereby removing any consistent temporal structure, recomputed the average LHL values, and calculated the same variability statistic. Repeating this procedure 10,000 times yielded a distribution against which we compared the observed test statistic. The resulting p-value was less than 0.0001, indicating that the observed LHL contour is highly unlikely to have occurred by chance and reflects a statistically significant pattern in the use of syncopation within a strain.

We then analyzed LHL contour over strains grouped by composer; results are shown in Figure 2. This graph suggests that the big three ragtime composers employed syncopation in distinct ways. To test this hypothesis, we conducted three pairwise one-way multivariate analyses of variance (MANOVA), comparing the mean LHL contours — treated as 16-dimensional vectors — between each pair of composers. In each test, the null hypothesis stated that the two composers had equivalent average syncopation contours. All three comparisons were statistically significant ($p < 0.0001$) after applying the Šidák correction, indicating that each composer employed a distinct pattern of syncopation within strains. These results support the hypothesis that the big three ragtime composers not only used syncopation with differing intensities, but also with structurally distinct temporal profiles.

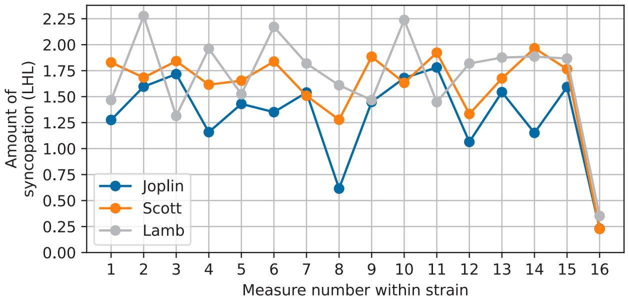

Figure 2. Average syncopation values during each measure of a 16-measure strain, grouped by composer.

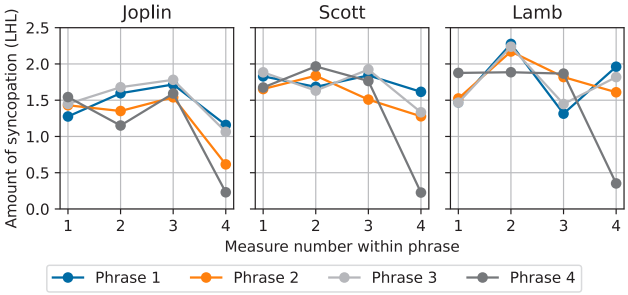

Figure 3. Average syncopation values during each measure of a 4-measure phrase within a strain.

It is also informative to examine syncopation use within the four-measure phrases of each strain. Figure 3 illustrates this by averaging LHL syncopation values across each four-measure phrase (i.e., measures 1–4, 5–8, 9–12, and 13–16) allowing for clearer identification of within-phrase trends. A consistent pattern emerges for Joplin and Scott: syncopation levels remain relatively stable during the first three measures of each phrase and then decrease noticeably in the fourth measure. Lamb, however, departs from this pattern. His phrases exhibit greater internal variability in LHL values, and the expected drop in syncopation at the end of the phrase occurs reliably only in phrases 2 (only slightly) and 4.

Our final analysis of syncopation patterns examines how syncopation changes across different strain categories. Most of the 71 compositions in the dataset contain four distinct strains, though a handful only have three, and a few have five or six. We examined average LHL value across the 16 measures of A, B, C, and D strains, using only data from the first variation of each strain. Figure 4 illustrates how A, B, and C strains tend to have relatively similar contours (the thinner lines) but D strains have a unique contour: these strains tend to alternate between more extreme high and low values of syncopation than the other strains. We ran another MANOVA test to examine if the contour of the D strains differed in a statistically significant way from the average contours of the A, B, and C strains *combined*. To get enough statistical power for the test, we used data from the first and second variations of each strain, rather than just the first. This test revealed a statistically significant difference in contour profiles between D strains and

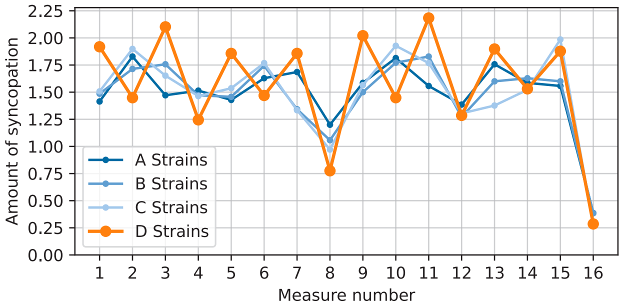

Figure 4. Average syncopation values during each measure of a 16-measure strain, grouped by strain identifier.

the other strains ($p < 0.01$).

4. PREDICTING PHRASE BOUNDARIES WITH MACHINE LEARNING MODELS

Having established that the three major ragtime composers — Joplin, Scott, and Lamb — use syncopation in measurably distinct ways within ragtime strains, we investigate whether these rhythmic differences can be leveraged to identify the boundaries of those same individual strains. This task falls under the broader umbrella of *phrase boundary detection*, a long-standing problem in computational musicology that has been studied extensively using both symbolic scores and audio recordings as input. Automated boundary detection has practical implications for music informatics, including interactive notation systems, structural visualization, and digital score analysis.

Previous Work. Approaches in the symbolic music domain typically fall into one of two categories: prescriptive systems usually based on rules derived from music theory, human perception, and cognition; and systems where rules or models are directly computed from data. Prescriptive systems include those found in Lerdahl and Jackendoff's *A Generative Theory of Tonal Music* [17], Narmour's Implication-Realization Model [18, 19], Cambouropoulos' Local Boundary Detection Model [20], and Temperley's Grouper model [21]. These systems attempt to encode human intuitions about melodic expectation and rhythmic grouping. While conceptually rich, many of these frameworks were not originally designed for computational implementation, though subsequent work has formalized parts of them algorithmically.

On the data-driven side, most techniques for phrase boundary prediction in symbolic data are based on statistical or machine learning approaches, including data-oriented parsing [22], information-theory-driven learning [23], restricted Boltzmann machines [24], rule mining approaches [25], and mathematical optimization [26]. In particular, we build on the investigations into neural networks for boundary prediction [27, 28] that examined convolutional neural networks (CNNs) and bidirectional long short-term memory networks (BiLSTMs); we extend this to include transformers and the attention mechanism [29].

Our Approach. We follow the symbolic data-driven paradigm, but this work differs in a few key areas. First, most prior work in this area, while acknowledging that much of phrase-boundary finding is rhythmically driven [25], relies on auxiliary melodic or harmonic information for an extra boost. In contrast, our goal was to examine what was possible with the rhythmic aspect of music notation alone, in particular, the binary onset pattern of each measure and its LHL syncopation value. Second, we use a novel dataset, and focus on predicting the start and/or end of phrase boundaries at the measure level, rather than the note level. We did this as to better align with perceptions of ragtime melodies which are tightly coupled to the 16-measure musical strain. Third, we build on prior machine learning models but focus on those designed for long-term dependency modeling, namely bidirectional long short-term memory networks (BiLSTMs) and transformers. Fourth, we expressly desired to test if it is easier to predict phrase boundaries for certain composers than others. To our knowledge, this is the first study to model phrase boundaries in symbolic ragtime scores using machine learning models with purely rhythmic features.

Task Formulation. We frame phrase boundary detection as a sequence labeling task. Each input is a ragtime composition, divided at the measure level. Each melody measure is represented by nine features: eight binary for the eight ones and zeros in the binary onset pattern, and one integer for the LHL value of the measure. All segment boundaries — both 16-measure ragtime strains and separate phrases not part of any strain — are encoded at the measure level using a 1 for the starting measure, 2 for the ending measure, and 0 for all other measures. Our goal was to develop a machine learning architecture that could predict the segment boundaries of an entire composition at once by labeling each measure of a previously-unseen composition with a 0, 1, or 2. We studied three different variations of labeling and predicting: using all three labels (0, 1, and 2), using only starting and continuing labels, and using only continuing and ending labels.

We evaluated eight machine learning models: two non-neural models to obtain a baseline level of accuracy, and six neural network architectures to evaluate the possible improvements that could be leveraged by models designed for capturing long-term dependencies in sequences. Our models are:

- **Logistic Regression**: The first baseline model.
- **SVM**: The second baseline model: a support vector machine with a radial basis function kernel to help capture non-linear decision boundaries.
- **CNN-CRF**: A convolutional neural network (CNN) architecture followed by a conditional random field (CRF) layer for sequence labeling. The CNN consists of a single 1D convolutional layer with 32 filters, a kernel size of 3, and ReLU activation, followed by a dropout layer with a rate of 0.3. The output of the convolution is passed through a fully connected layer that maps to the label space, producing per-measure class scores. Finally, a CRF layer, inspired by [28], is used to model transitions between measure labels and enforce structural consistency in the predicted sequence. The total number of parameters is 1930.

- **CNN-3-CRF**: An extended version of the previous model with a deeper architecture consisting of three stacked convolutional layers, each similar to the single layer in the previous model: 32 filters, a kernel size of 3, and ReLU activation. The final output passes through a dropout layer with a rate of 0.3, followed by a fully connected layer identical to that in the previous model. This deeper CNN captures increasingly abstract rhythmic features over a larger temporal context, with the goal of identifying more complex segment boundary cues. The total number of parameters is 7178.
- **BiLSTM**: A bidirectional LSTM (BiLSTM) model with two stacked layers, each with 32 hidden units per direction (yielding 64-dimensional outputs per time step after concatenation). A dropout layer with a rate of 0.3 is applied between layers. The LSTM outputs are passed through two fully connected layers: a linear projection from 64 to 32 dimensions, followed by a ReLU activation and dropout, and a final linear layer mapping from 32 to the number of label classes. The BiLSTM architecture is well-suited to this task because it captures both past and future context for each time step, making it effective at detecting temporal patterns such as rhythmic transitions and segment boundaries that depend on information across the sequence. The total number of parameters is 38,242.
- **BiLSTM-CRF**: A BiLSTM identical to the previous one, but with a CRF layer applied on top. The total number of parameters is 38,250.
- **Transformer**: A transformer model that first projects each measure of music into a 128-dimensional embedding space using a linear projection layer. A positional encoding is then added to preserve the order of measures. The embedded sequence is processed by a stack of two Transformer encoder layers, each with four attention heads and an embedding size of 128. The output is passed through two fully connected layers: first from 128 to 128 hidden units with ReLU activation and dropout with a rate of 0.3, and then from 128 to the number of output label classes. Unlike the BiLSTM models, which process sequences sequentially and capture temporal dependencies in order, the transformer attends to all positions simultaneously, allowing it to model long-range rhythmic dependencies more efficiently and capture global patterns in the music. The total number of parameters is 36,834.
- **Transformer-CRF**: A transformer identical to the previous one, but with a CRF layer applied on top. The total number of parameters is 36,842.

Each model was coded in PyTorch [30] or scikit-learn [31] and trained using the Adam optimizer with a weight decay of 0.00005 for L2 regularization, and early stopping. Hyperparameters (number of CNN filters, BiLSTM and transformer dimensions) were tuned separately on a fixed 20% subset of the data and held constant during leave-one-out cross validation. We used a loss function adjusted for class imbalance as appropriate (cross-entropy or negative log-likelihood), depending on model. We evaluated each model under three phrase boundary prediction setups (predicting start-continue-end, start-continue, and continue-end), two dataset variations (whole pieces labeled versus only 16-measure strains being labeled), and four composer training/testing groupings (all composers, only Joplin, only Scott, and only Lamb). For each of these setups, we ran leave-one-out cross validation at the composition level, predicted phrase boundaries for each measure of the tested piece, and calculated precision, accuracy, and F1 score. For every configuration, we also predicted segment boundaries with the true LHL input ablated (set to zero in the input) to simulate training each model with only note onsets and no LHL values.

Results. On average, across all task variations, the models trained to locate the end of segments (trained on segments labeled with 0 and 2) had higher average F1 scores than the models trained to locate segment beginnings (labels 0 and 1) or both (all three labels). This was true for 7 out of 8 models, both dataset variations, and all composer groups. Therefore, to save space, we present results here only for the end-of-segment labeling task; full results for the other labeling tasks are available in the online code repository. Table 1 shows precision, recall, and F1 scores for all task combinations predicting segment endings; because this task is trained to label the measures of a composition with 0 for a segment continuation and a 2 for an ending, showing just the results for the ending predictions is sufficient and makes it easier to compare our results with other studies; recent studies cite F1 scores ranging from 0.26 to 0.89 [24, 25, 28].

Including the LHL syncopation feature in the inputs improved performance across all model types, training subsets, and composer groups, with an average F1 score increase of 0.02. While modest, this gain is statistically significant ($p < 0.0001$) according to a Wilcoxon signed-rank test. One likely explanation for this increase is that the LHL metric is derived directly from the same binary onset patterns already available to the models, dampening its effect. In future work, we plan to explore whether LHL values, especially when calculated over longer contexts than one measure, interact with specific metrical subsets of binary onset pattern values in ways that support more accurate boundary detection.

On average, models performed slightly better when trained on all segment types (i.e., both 16-measure strains and miscellaneous phrases such as introductions and codas), likely because strain and non-strain boundaries share rhythmic characteristics, providing a broader distribution of boundary patterns. We isolated results from this setting in Table 1 and visualized them in Figure 5 to highlight several key findings.

First, the neural network models consistently outperformed the logistic regression and SVM baseline models. While this may not be surprising given their greater number of trainable parameters, it is notable because the statistical patterns described in Section 3 might suggest that this task is relatively straightforward. In practice, however, identifying phrase boundaries requires more nuanced modeling than simple classification can provide. The CNN-based models showed modest gains over the baselines, but the

Model	Training	All composers Pre	Rec	F1	Δ	Only Joplin Pre	Rec	F1	Δ	Only Scott Pre	Rec	F1	Δ	Only Lamb Pre	Rec	F1	Δ
Logistic Reg	All	0.23	0.87	0.36	+0.01	0.19	0.87	0.31	+0.02	0.29	0.89	0.44	-0.01	0.22	0.83	0.35	+0.01
SVM	All	0.27	0.91	0.41	+0.00	0.25	0.88	0.38	+0.00	0.37	0.91	0.52	+0.00	0.21	0.80	0.34	+0.00
CNN-CRF	All	0.56	0.43	0.49	+0.04	0.50	0.32	0.39	+0.03	0.55	0.56	0.55	+0.11	0.54	0.43	0.48	+0.06
CNN-3-CRF	All	0.48	0.55	0.52	+0.02	0.44	0.43	0.43	+0.04	0.53	0.61	0.57	+0.07	0.57	0.53	0.55	+0.00
BiLSTM	All	0.56	0.80	0.66	+0.03	0.45	0.57	0.50	+0.01	0.61	0.81	0.70	+0.01	0.51	0.73	0.60	-0.02
BiLSTM-CRF	All	0.76	0.70	**0.73**	+0.05	0.54	0.46	0.50	+0.02	0.81	0.72	**0.76**	+0.03	0.65	0.58	0.62	+0.04
Transformer	All	0.51	0.88	0.65	+0.00	0.54	0.79	0.64	-0.01	0.56	0.86	0.68	+0.00	0.54	0.82	**0.65**	+0.01
Transf-CRF	All	0.72	0.72	0.72	+0.03	0.71	0.69	**0.70**	+0.02	0.74	0.75	0.75	+0.05	0.62	0.67	0.64	+0.01
Logistic Reg	Only 16	0.19	0.85	0.31	+0.01	0.15	0.84	0.26	+0.01	0.24	0.88	0.37	+0.00	0.20	0.84	0.32	+0.01
SVM	Only 16	0.22	0.91	0.36	+0.00	0.20	0.87	0.32	+0.00	0.31	0.89	0.46	+0.01	0.20	0.76	0.32	-0.01
CNN-CRF	Only 16	0.55	0.36	0.43	+0.05	0.50	0.31	0.39	+0.08	0.50	0.47	0.48	+0.10	0.49	0.35	0.41	+0.03
CNN-3-CRF	Only 16	0.44	0.51	0.47	+0.05	0.43	0.43	0.43	+0.05	0.50	0.56	0.53	+0.15	0.47	0.33	0.39	-0.04
BiLSTM	Only 16	0.54	0.78	0.64	+0.02	0.45	0.65	0.53	+0.01	0.58	0.77	0.66	+0.03	0.47	0.68	0.55	+0.01
BiLSTM-CRF	Only 16	0.75	0.67	0.70	+0.03	0.60	0.52	0.56	+0.09	0.73	0.63	0.68	+0.03	0.67	0.55	0.61	+0.03
Transformer	Only 16	0.45	0.88	0.59	-0.01	0.44	0.78	0.56	-0.02	0.48	0.89	0.62	+0.00	0.46	0.75	0.57	+0.01
Transf-CRF	Only 16	0.75	0.70	**0.73**	+0.06	0.69	0.67	**0.68**	+0.03	0.68	0.71	**0.69**	+0.05	0.61	0.66	**0.63**	+0.01

Table 1. Performance on predicting segment boundaries by model, subset of data used (all segments or only 16-measure strains), and composer. Δ indicates improvement in F1 score when the LHL syncopation value is included in training.

BiLSTM and transformer models achieved around double the F1 scores of the simpler models across most test conditions. The CNN models likely lag behind because while they can capture local rhythmic patterns in the measures well, they lack a mechanism for modeling long-range dependencies, an essential feature for identifying phrase structure. These results indicate that the deeper architecture in the 3-layer CNN as compared to the single-layer CNN helped performance only slightly, even with an almost four-fold parameter increase. In contrast, the BiLSTM and transformer models can more easily model temporal dependencies across larger time spans.

Second, we note that in comparing performance in all composers grouped together, the CRF and non-CRF versions of the BiLSTM and transformer models performed similarly, with the CRF giving a slight boost to the F1 scores, most likely due to the CRF's capacity to enforce structured label transitions. This is noteworthy as these models have similar numbers of trainable parameters.

Third, model performance varies notably by composer. Even though Joplin and Scott each had roughly the same numbers of pieces in the corpus (31 versus 28), finding phrase boundaries in Joplin's compositions appeared to be a harder task for six of the eight models; only the transformer-based classifiers seemed to cope roughly equally well with the two composers' phrases. Lamb, with only 12 pieces, exhibited the greatest relative variability in F1 scores, sometimes being the most difficult of the three composers and sometimes being almost even with Joplin and Scott. These differences likely stem from distinct rhythmic styles: as shown in Figure 3, Joplin's strains exhibit frequent reductions in syncopation at the end of each four-measure phrase, which may lead the model to misinterpret internal boundaries as final ones. By contrast, Scott and Lamb tend to reserve syncopation drop-offs for the end of full 16-measure strains, thereby providing clearer predictions for segment boundaries. The sparsity of Lamb's data likely further contributes to inconsistency in model performance.

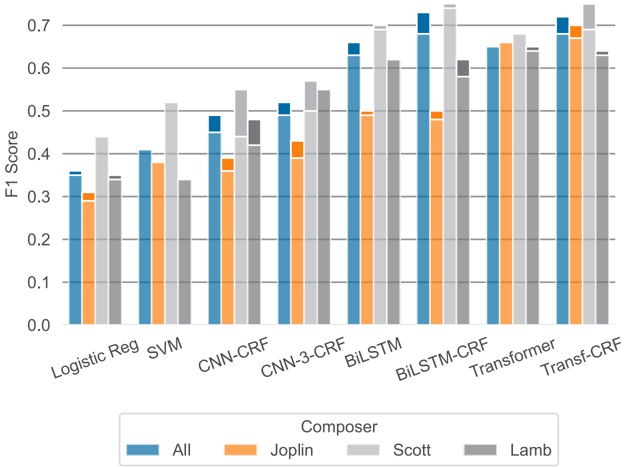

Figure 5. Performance of each model and composer group in predicting strain endings. Improvement due to including LHL is shown by the shaded region at the top of applicable bars.

These findings demonstrate that rhythmic notation alone encodes sufficient information to identify phrase boundaries, even in the absence of pitch or harmonic content. Additionally, the differences in syncopation and phrase shaping among the composers in the corpus are reflected in the models' performance. Further musicological investigations could explore whether this rhythm-based approach generalizes to other metrically strong musical genres, potentially suggesting insights for both symbolic music representation and the analysis of historical compositional style.

5. PREDICTING COMPOSERS

While the previous section focused on how rhythmic features such as syncopation align with phrase structure in ragtime, we now turn to the question of whether these features can also be used to distinguish between individual composers. In other words, we ask if Joplin, Scott, and Lamb employ rhythm and syncopation in ways that are compositionally distinct enough to allow machine learning models to identify authorship based on notated rhythms alone.

Previous Work and Task Formulation. Composer attribution has been a longstanding task in music informatics. Traditional approaches often relied on hand-crafted melodic, harmonic, or rhythmic features, but more recent deep learning methods have enabled direct classification from symbolic scores [32, 33, 34]. Our approach straddles the line: as in the last section, our goal is to see what is possible using solely a rhythmic view of a score. Therefore, we use similar machine learning models as in the previous section, that are provided only binary onset patterns and syncopation values. We evaluate two task formulations: (1) strain-level classification, where each 16-measure segment is independently labeled and predicted, and (2) piece-level classification, where the entire composition is input and a single composer label is predicted.

We adapted five of the eight models from the previous section to work in this new formulation. Specifically, we used the logistic regression, SVM, 3-layer CNN, BiLSTM, and transformer models. We removed the CRF layer from the CNN (and did not use it in the other neural network models) because it was unsuitable for this task: a conditional random field is specifically used in situations with multiple labels being predicted to model dependencies between them; here, there is only one label per prediction (a single 16-measure strain or composition). For this same reason, each neural network model also applies pooling after processing its input. The CNN models use global average pooling to summarize convolutional activations over time, the BiLSTM model uses mean pooling over the sequence dimension, and the transformer uses mean pooling across the time dimension. These operations are all mathematically identical, though the terminology used by convention is slightly different. As in the previous section, we incorporated class weights into the cross-entropy loss function.

For predicting composer from a strain, we used 10-fold cross validation, and for predicting composer from an entire piece, we used leave-one-out cross-validation. Results are presented in Table 2, with confusion matrices in Figure 6. Direct comparison to prior work is difficult due to differences in dataset composition and input modality; most notably, our models rely exclusively on rhythmic features. However, for reference, a similar three-composer prediction task in [34] cited per-class accuracies rates between 0.864 and 0.998 (compare with our recall rates).

Results. Several insights emerge from these experiments. First, strain-level models generally outperform whole-piece models across the board. This is likely due to the greater number of training examples available at the segment level,

		Strain-level			Piece-level		
Model	Comp	Pre	Rec	F1	Pre	Rec	F1
Logistic Reg	Joplin	0.78	0.69	0.73	0.73	0.77	0.75
Logistic Reg	Scott	0.71	0.77	0.74	0.63	0.61	0.62
Logistic Reg	Lamb	0.51	0.56	0.54	0.36	0.33	0.35
SVM	Joplin	0.94	0.88	0.90	0.67	0.90	0.77
SVM	Scott	0.86	0.94	0.90	0.62	0.64	0.63
SVM	Lamb	0.92	0.88	0.90	0.00	0.00	0.00
CNN-3	Joplin	0.80	0.78	0.79	0.67	0.71	0.69
CNN-3	Scott	0.73	0.74	0.73	0.64	0.57	0.60
CNN-3	Lamb	0.50	0.50	0.50	0.23	0.25	0.24
BiLSTM	Joplin	0.75	0.80	0.77	0.67	0.65	0.66
BiLSTM	Scott	0.72	0.73	0.72	0.63	0.68	0.66
BiLSTM	Lamb	0.50	0.39	0.44	0.27	0.25	0.26
Transformer	Joplin	0.81	0.83	0.82	0.76	0.84	0.80
Transformer	Scott	0.75	0.74	0.74	0.60	0.54	0.57
Transformer	Lamb	0.52	0.52	0.52	0.25	0.25	0.25

Table 2. Per-class performance on predicting composers from individual strains or complete pieces.

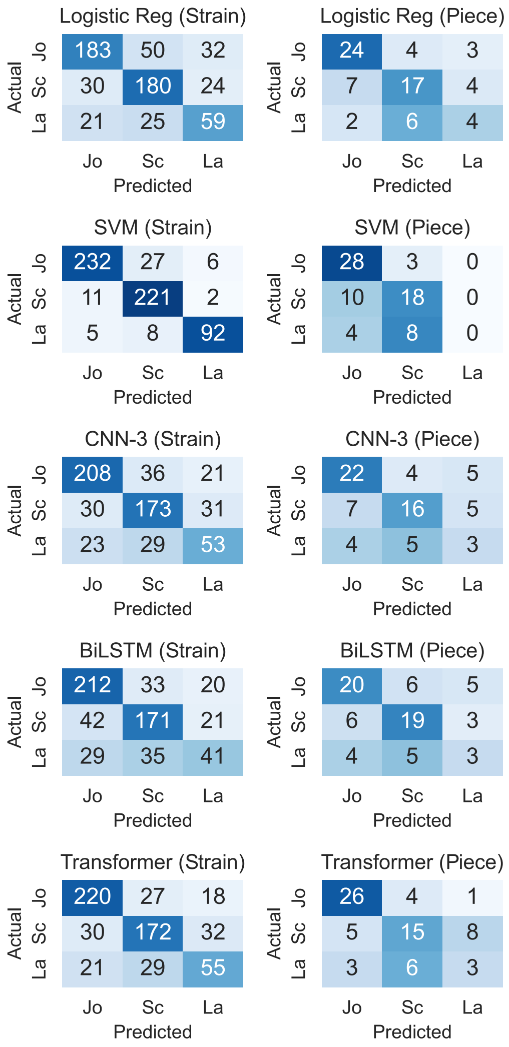

Figure 6. Confusion matrices for strain-level and piece-level classification tasks.

as well as the models' effectiveness at detecting short-term rhythmic patterns. Second, the SVM achieves the highest strain-level F1 scores across all composers, suggesting that simpler models can excel when the prediction context is constrained. In contrast, the SVM failed to classify any Lamb compositions correctly in the piece-level task, and continued to misclassify such compositions even when we oversampled Lamb scores in the training data. Lamb was consistently the most difficult composer to classify for each model, which may reflect both the smaller number of available pieces (12) and higher within-class variability. At the piece level, Lamb was frequently misclassified as Scott, which may stem from the fact that both composers exhibit higher average syncopation than Joplin.

Third, these two tasks see less improvement by using neural network models over the baseline models (logistic regression and SVMs) than in the segment boundary detection task from the previous section. Here, in the the strain-level classification task, the three neural network models show only a slight improvement over logistic regression and none at all over the dominantly-performing SVM. In the piece-level classification task, the neural models actually perform slightly worse, on average, than logistic regression and the SVM. This is somewhat surprising, as one would expect that these architectures, especially the BiLSTM and transformer, would do better at capturing higher-level rhythmic and temporal information in the music; however, this suggests that in ragtime, composer identity is primarily expressed through local rhythmic motives rather than extended structural dependencies.

Several other factors may explain these results. First, though we saw in Section 3 (Figures 2 and 3) that there are statistically-significant differences in the way that the three composers use syncopation, these differences are relatively subtle. Second, the dataset is small, which limits the potential of higher-capacity models and increases the risk of overfitting. With these two factors, we might have reached the limit of how well the discussed machine learning models can do.

Musicological Analysis. To further interpret these results through a musicological lens, we examined the weights that logistic regression achieved on the strain-level composer classification task; these weights are illustrated in Figure 7 on a per-composer basis. Because the features are normalized, we can interpret the weights as representing the contribution of each feature toward predicting a given composer, relative to the others. Each of the first seven features corresponds to a binary onset at a specific metric position in the measure ("Pos0" through "Pos7"); these are followed by the LHL syncopation value feature and a bias term. Because each feature corresponds directly to a position in the notated measure, we obtain a clear mapping from the computational output back to the original symbolic rhythmic notation.

The weight profiles suggest clear stylistic differences in rhythm and syncopation usage across composers. We can make a few observations. As expected from our earlier statistical analyses (Figures 2 and 3), Scott and Lamb receive higher weights on the LHL syncopation feature than Joplin, reflecting their greater use of syncopation. More granular insights can be derived from the individual positional weights. For instance, while the downbeat of the measure, position 0, does not show much difference between the composers, the halfway point, position 4, shows large differences between Joplin and Scott: Joplin appears to favor having an onset here much more than Scott does, suggesting that Scott included more syncopation crossing the midpoint of the measure (e.g., from position 3 to position 4), than Joplin did. The opposite situation occurs at position 6, where an onset here favors Scott but not Joplin, suggesting Joplin preferred syncopation crossing from position 5 to 6 more than Scott did. The large negative bias weight for Lamb reflects his underrepresentation in the corpus.

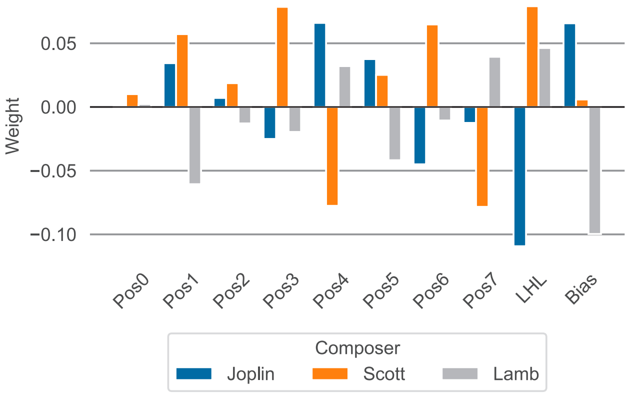

Figure 7. Logistic regression weights for each feature for strain-level composer classification. Positive weights indicate that higher feature values increase the likelihood of classifying a segment as that composer relative to the other two composers, while negative weights reduce it.

These weight differences reinforce a musicological interpretation of stylistic rhythm in ragtime: while all three composers make use of syncopation, the specific rhythmic patterns and degrees of metric disruption vary in ways that are detectable and classifiable. Though the precision and recall scores that logistic regression obtained in this experiment limit the overall power of these musicological interpretations, it is still a useful tool.

These results and their musicological interpretation suggest several avenues for future work. From a modeling perspective, hybrid architectures that combine convolutional layers for local rhythmic motif detection with recurrent or attention-based layers for long-range structure may better capture both local and global rhythmic features. Expanding the dataset, particularly with additional Lamb compositions, could reduce class imbalance and when combined with an interpretable model, would provide more opportunities for useful musicological findings. Finally, it remains an open question whether the rhythmic dimension of music notation alone is sufficient for composer attribution in this repertoire, or whether additional melodic or harmonic information is needed to more fully distinguish compositional style.

6. CONCLUSION

This paper demonstrates that the rhythmic aspect of music notation alone, represented by binary onset patterns and Longuet-Higgins and Lee (LHL) syncopation values, provides substantial information about both musical structure and composer identity in ragtime. Using only these symbolic rhythmic representations, we showed that machine learning models, particularly those like BiLSTMs or transformers that are designed to detect long-term structural dependencies, can reliably identify the boundaries of ragtime strains. On the composer-classification task, non-neural models approached or exceeded the performance of neural networks. In particular, support vector machines performed best in the strain-level task, but were inconsistent in the piece-level task. Still, the overall performance statistics indicate that rhythmic notation provides a good amount of guidance for distinguishing composers, especially those of Joplin and Scott.

From a musicological perspective, these findings reinforce that rhythmic patterns encode musical style and structural information in a way that is both detectable and classifiable. The interpretable logistic regression weights and positional feature analysis highlight how differences in phrasing and syncopation among Joplin, Scott, and Lamb are evident directly in their notated rhythmic patterns. This underscores the value of score-based rhythmic representations for both computational musicology and the study of historical musical style.

Future work will explore several directions. Modeling segment-level syncopation trajectories, rather than per-measure values, may improve boundary detection and composer attribution. Hybrid neural network architectures that combine convolutional layers for local rhythmic motifs with attention-based or recurrent layers for global context could further improve performance. Expanding the dataset could enable more robust stylistic modeling and support deeper musicological insight. Finally, extending this work to evaluate the contributions of melodic and harmonic features alongside rhythmic notation will clarify how different aspects of score representation interact in composer attribution and segment boundary detection.

7. REFERENCES

[1] E. A. Berlin, "Ragtime," in *Grove Music Online*. Oxford University Press, 2013. [Online]. Available: https://doi.org/10.1093/gmo/9781561592630.article.A2252241

[2] R. Blesh and H. Janis, *They All Played Ragtime*, 4th ed. New York: Oak Publications, 1971.

[3] H. C. Longuet-Higgins and C. S. Lee, "The rhythmic interpretation of monophonic music," *Music Perception*, vol. 1, no. 4, pp. 424–441, 1984.

[4] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press, 2006.

[5] G. Sioros, G. Madison, D. Cocharro, A. Danielsen, and F. Gouyon, "Syncopation and groove in polyphonic music: Patterns matter," *Music Perception*, vol. 39, no. 5, pp. 503–531, 2022.

[6] J. E. Hasse, "Ragtime: From the top," in *Ragtime: Its History, Composers, and Music*, J. E. Hasse, Ed. New York: Shirmer Books, 1985, pp. 1–39.

[7] I. Harer, "Defining ragtime music: Historical and typological research," *Studia Musicologica Academiae Scientiarum Hungaricae*, vol. 38, no. 3–4, pp. 409–415, 1997.

[8] A. Volk and W. B. de Haas, "A corpus-based study on ragtime syncopation," in *Proc. of the 14th International Society for Music Information Retrieval Conference*, 2013, pp. 163–168.

[9] P. B. Kirlin, "A corpus-based analysis of syncopated patterns in ragtime," in *Proc. of the 21st International Society for Music Information Retrieval Conference*, 2020, pp. 647–653.

[10] H. V. Koops, A. Volk, and W. B. de Haas, "Corpus-based rhythmic pattern analysis of ragtime syncopation," in *Proc. of the 16th International Society for Music Information Retrieval Conference*, 2015, pp. 483–489.

[11] J. Yust and P. B. Kirlin, "The multileveled rhythmic structure of ragtime," in *Proc. of the 9th International Conference on Culture and Computing*, 2021, pp. 337–354.

[12] D. Bemman and J. Christensen, "Inner metric analysis as a measure of rhythmic syncopation," in *Proc. of the 25th International Society for Music Information Retrieval Conference*, 2024, pp. 389–396.

[13] A. Volk, "The study of syncopation using inner metric analysis: Linking theoretical and experimental analysis of metre in music," vol. 37, no. 4, pp. 259–273, 2008.

[14] G. Sioros and C. Guedes, "Complexity driven recombination of MIDI loops," in *Proc. of the 12th International Society for Music Information Retrieval Conference*, 2011, pp. 381–386.

[15] G. Sioros, A. Holzapfel, and C. Guedes, "On measuring syncopation to drive an interactive music system," in *Proc. of the 13th International Society for Music Information Retrieval Conference*, 2012, pp. 283–288.

[16] D. A. Jasen and T. J. Tichenor, *Rags and Ragtime: A Musical History*. New York: Seabury Press, 1978.

[17] F. Lerdahl and R. Jackendoff, *A generative theory of tonal music*. Cambridge, MA: MIT Press, 1983.

[18] E. Narmour, *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. Chicago: University of Chicago Press, 1990.

[19] ——, *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model*. Chicago: University of Chicago Press, 1992.

[20] E. Cambouropoulos, "Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface," in *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology*, M. Leman, Ed. Springer-Verlag, 1997, pp. 388–417.

[21] D. Temperley, *The Cognition of Basic Musical Structures*. Cambridge, MA: MIT Press, 2004.

[22] R. Bod, "Memory-based models of melodic analysis: Challenging the gestalt principles," *Journal of New Music Research*, vol. 31, no. 1, pp. 27–36, 2002.

[23] M. T. Pearce, D. Müllensiefen, and G. A. Wiggins, *Melodic Grouping in Music Information Retrieval: New Methods and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 364–388.

[24] S. Lattner, C. E. C. Chacón, and M. Grachten, "Pseudo-supervised training improves unsupervised melody segmentation," in *Proc. of the 24th International Joint Conference on Artificial Intelligence*, 2015, pp. 2459–2465.

[25] P. van Kranenburg, "Rule mining for local boundary detection in melodies," in *In Proc. of the 21st International Society for Music Information Retrieval Conference*, 2020, pp. 271–278.

[26] M. E. R. López, "Automatic melody segmentation," Ph.D. dissertation, Utrecht University, Utrecht, 2016.

[27] K. Ullrich, J. Schlüter, and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in *Proc. of the 15th International Society for Music Information Retrieval Conference*, 2014.

[28] Y. Zhang and G. Xia, "Symbolic melody phrase segmentation using neural network with conditional random field," in *Proc. of the 8th Conference on Sound and Music Technology*, 2020, pp. 55–65.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8024–8035.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] G. Buzzanca, "A supervised learning approach to musical style recognition," in *Proc. of the International Conference on Music and Artificial Intelligence*, 2002.

[33] G. Velarde, T. Weyde, C. C. Chacón, D. Meredith, and M. Grachten, "Composer recognition based on 2D-filtered piano-rolls," in *Proc. of the International Society for Music Information Retrieval Conference*, 2016, pp. 115–121.

[34] H. Verma and J. Thickstun, "Convolutional composer classification," in *Proc. of the 20th International Society for Music Information Retrieval Conference*, 2019, pp. 549–556.