Automatic Self-Similarity Based Form Labelling of Classical-Period Piano Sonata Movements From Audio Recordings

Paul A.D. Burger
Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, South Africa
CSIR, Pretoria, South Africa

J. Pieter Jacobs
Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, South Africa

Abstract:

Musical form refers to the overall structure or organisation of a musical composition. It is a complex and high-level property of music that requires musical training to identify. A review of previous research in this field indicates that the focus has been on the task of detecting section boundaries and that automatic audio based form label recognition is a field of study that remains largely unexplored. This study explores the complex task of automatically determining musical form from audio. It demonstrates the ability of a novel methodology to label eight different form types that occur in the movements of Classical-period piano sonatas. The methodology makes use of self-similarity matrices, generated from features extracted from raw audio, as input to a convolutional neural network. The superiority of our approach was confirmed by evaluating it against a neural network model based on state-of-the-art features. We also report an evaluation of self-similarity matrices based on automatically transcribed piano rolls for the task of form recognition. Piano rolls are demonstrated to be superior for this application when compared to a range of other feature representations. Additionally, the performance of the model is shown to be robust in handling variations in performer choices. These range from different interpretations of the same score to actual deviations from the score where performers may elect to play or not to play notated repeats thus highlighting its ability to generalise across different performances of the same piece.

I Introduction

Research in the field of music information retrieval (MIR) focuses largely on the task of analysing audio content for the purposes of music categorisation and retrieval. Being able to describe audio content in terms of musically meaningful concepts is an important facet of MIR [1]. The task of MIR is to extract relevant musical features, and to make use of these features for the purposes of indexing and the development of algorithms that can be used to search and retrieve musical content that is relevant to users [2].

Musical form refers to the overall structure of a piece of music which results from the impulse to create order and structure [3]. Form refers to the largest shape of a piece of music. It is the result of the interaction between melody, harmony, rhythm, and timbre [4]. It guides a listener through the piece by balancing contrast and repetition in different ways. Musical form is a valuable high-level feature which can enhance music recommendation and retrieval systems by helping them identify pieces that are relevant to listeners. Most genres of music exhibit formal properties; for example, jazz makes use of standard forms.

Automatic music structure analysis that takes audio recordings as input can take place on two levels, namely overall form labelling and detailed segmentation [5] (for the sake of conciseness, we omit the adjective “overall” in what follows). The goal of detailed audio segmentation is to identify the exact times in a music recording where one segment ends and the other starts, and can be used to provide snippets of audio (such as the chorus of a song) to users browsing online music listening services, for example. Automatic form labelling, on the other hand, assigns labels regarding the structure or form of an entire piece of music. Audio segmentation has been well covered in the literature [5], while form labelling remains an understudied problem; we are aware of one other study that attempts to directly produce form annotations for classical audio recordings [6]. A possible reason for this is a shortage of datasets containing high-level form annotations. Creating appropriate datasets for form recognition is invariably more expensive than datasets for section boundary detection, because in the former case each piece only provides one sample to the dataset, whereas a single piece can have multiple segment boundaries. Another factor is the difficulty of the label

annotation task. In practice, compositions can, and often do, deviate from the rigorous definitions of the various forms. Based on their independent analyses, musicologists may therefore reach differing conclusions regarding the form of a piece [7], [8], [9].

In the present work, which expands on [10], we take a direct approach to automatic form labelling that circumvents a detailed segmentation stage. With a bottom-up hierarchical analysis approach that relies on detailed segmentation, it may be unclear whether the identified segment boundaries indicate transitions between larger sections or subsections. Furthermore, it can be challenging to identify the relevant musical property defining a boundary at any point in time, as boundaries may be based on harmonic information, timbre, and/or tempo [11]. Once segments have been identified, it may still not be straightforward to identify the overall form type from them.

Specifically, we introduce a methodology for identifying eight fundamental form types found in Classical-period piano sonata movements from audio recordings of these movements (this period is of enormous importance in Western music history). The basic idea is to present self-similarity matrices (SSMs) of the full movements as input to a convolutional neural network (CNN). These matrices implicitly encode the structure of the music piece as a two-dimensional matrix. The SSMs are constructed from spectrogram-based features derived from the raw audio on the one hand, and symbolic (i.e., note-level) features obtained from a state-of-the-art neural network piano transcription system that takes spectrograms as input, on the other hand. We chose convolutional neural networks (CNNs) to transform these images into form labels given their proven capabilities with regard to image processing tasks. By directly generating form labels from the self-similarity matrices, we avoid the difficulty of inferring form labels in a bottom-up manner. We specifically focused on music for solo piano because of the current availability of good piano transcription algorithms, such as the above, that can be used towards generating SSMs based on piano roll features, which proved invaluable when determining the form label of a movement.

We developed a new custom dataset consisting of form labels for 480 movements from sonatas by five iconic composers from the Classical period. The labels were collected from several musicological resources. Instances were found where different form labels were assigned to the same movement from the same piece by different musicologists. For example, Tovey describes the fourth movement of Beethoven's Sonata No. 11 as having a rondo form [12], while Stainkamph characterises it as having a rondo sonata form [13] (a related form type). Consequently, we framed our study as a multi-label classification problem, rather than adopting the multi-class classification approach used by other researchers [6]. In multi-class classification the problem is defined so that the musical content may only belong to one category, whereas in multi-label classification problems, the musical content can belong to multiple classes simultaneously.

In addition to the above methodology and dataset, our contributions include the following:

- We demonstrate the superiority of our approach by evaluating it against a neural network model based on the current state-of-the-art [6], setting a benchmark for this task.
- We perform a detailed analysis of the predictive functioning of our model in the context of differing performances of different pianists of the same piece that are motivated by interpretative choices (e.g., related to expressive timing), and their adherence (or not) to notated repeats in the score. It turns out that our approach is robust to both types of variabilities.
- We confirm that the improvement in predictive performance achieved by CNNs trained on SSMs based on symbolic features (i.e. velocity piano rolls) vs. SSMs based on features directly derived from mel-spectrograms is statistically significant.

This article is organised as follows: Section II provides an overview of musical form, while Section III contains a literature review. In Section IV, the custom dataset created for this study is described. Section V outlines the methodology used, and Section VI presents the findings. Finally, Section VII discusses the insights gained and the conclusions drawn.

II Musical Form

As noted above, the largest shape of a piece of music is its form; it results from the interaction between melody, harmony, rhythm, and timbre [4]. In practice, form is generally identified by subdividing the piece into smaller distinct units called phrases. A musical phrase expresses some musical thought or idea but usually lacks the weight to stand alone [14]. These phrases are varied and/or combined with other phrases to form a section. The organisation of sections defines the form of the piece. These sections are usually designated with letters when being analysed. For example, the first section is given the letter "A." The following contrasting section would be designated "B." If the material from the first section appears elsewhere in the piece, it will also be designated with the letter "A." Sometimes larger sections are further broken down. Lowercase letters are used for repetition and contrast within larger sections. Repetition of sections is indicated by ||: and :||, where ||: marks the start of the section to be repeated, and :|| marks the end of the section to be repeated. When a variation of a section is played, the section letter appears with an accent, e.g., A'. The movements of Classical period piano sonatas in musicological sources [7], [8], [9], [12], [13], [15], [16], [17], [18], [19], [20], [21] are predominantly assigned to the following form types:

- *Binary:* A musical piece that has two sections of approximately equal duration. The first section is characterised by an articulated movement towards another key (usually the dominant one). In the second section, the piece returns to the tonic key [22]. Binary form is symbolised by the letters AB [15].
- *Ternary:* Based on the principle of departure and return and of thematic contrast and repetition, ternary form normally contains three distinct sections. The first section is always repeated in the third section and is generally unchanged, while the middle section highly contrasts the surrounding material. Even if the material from the first section is continued it will be changed in some substantial way, for example, mode, scoring or tempo [23]. Ternary form is symbolised by the letters ABA [15].
- *Sonata form:* Considered to be the most consequential form type of instrumental music during the Classical period [24], sonata form consists of three sections that are embedded in a two-part tonal structure. Sonatas start with the exposition, which forms the first section and the first part of the two-part tonal structure. The second part of the structure includes the development and the recapitulation [25]. Laying out the main thematic material for the movement, the exposition is often repeated. Themes in the exposition will be altered and varied in the development to create contrast. The recapitulation resolves the progression, completing the harmonic movement. Sonata form can be symbolised by the letters ABA' [15]. A more detailed breakdown of sonata form according to melodic organization is as follows [26]:

  Exposition Development Recapitulation: (PT Trans. ST CT) (FS Retransition) (PT Trans. ST CT), where PT is the primary theme, ST is the secondary theme, CT is a closing theme, and FS denotes fragmentation and sequencing.

- *Minuet/Scherzo and Trio:* A minuet is a French dance originally performed by the nobility. The dance is performed in a moderate or slow triple meter. It appears frequently in the movements of late 18th-century multi-movement forms where it was normally paired with a trio [27]. Scherzo comes from the German word *Scherz* ("joke"). The term has been used to describe any movement that takes the place of a minuet and trio in a sonata cycle. It is also used to indicate a piece that is comical or ironic and is usually fast-paced [28]. Scherzos are also usually paired with a trio. Minuet and trio and scherzo and trio are both specific types of ternary form [15] symbolised by:

$$\underset{\text{||:a:||}\ \text{||:ba}'\text{ : ||}}{\text{A}} \qquad \underset{\text{||:c:||}\ \text{||:dc:||}}{\text{B}} \qquad \underset{\text{aba}'}{\text{A}}$$

- *Rondo:* A multi-section movement where the main theme starts the piece and then recurs, normally in the home key, between subsidiary sections, before returning to the main theme to finish the piece [29]. Rondo form is symbolised by the letters ABACA for the small rondo and ABACAB ' A for the Classical rondo [15].

- *Rondo sonata:* A formal procedure that combines elements of the sonata with that of the rondo to produce a hybrid. In practice, this is achieved by uniting the tonal and recapitulatory arrangement characteristics of the sonata form with the characteristic rondo principle of returning to the initial idea [30]. Rondo sonata form can be thought of as being in three main parts, symbolised as part one: ABA; part two: C; and part three: ABA [15].
- *Theme and variations:* Theme and variations is another popular form that was used extensively. It is characterised by continuous variation based on a theme [15]. In theme and variations form, an introductory theme is stated which is then modified in every subsequent repetition. Rather than juxtaposing the contrast with the restatement (as in binary or ternary forms), contrast is merged with the restatement [15]. Theme and variations form can be symbolised by: AA ′ A′′A′′′ ...

III Related Work

Although there are datasets available for testing musical structure analysis algorithms, they often lack labels that describe the overall form of a piece. The Structural Analysis of Large Amounts of Music Information (SALAMI) [31] dataset, for example, provides section annotations for a wide range of classical and non-classical music, but interpreting the overall form of a piece from these annotations can be challenging. By contrast, the Standardized Musical Form and Structure Analysis (SMFSA) database [32] includes both form and section annotations specifically for classical music. This dataset consists of 200 pieces, with annotations created by human experts. It is diverse in terms of compositional style (including sonatas, concertos, requiems, etc.), musical period (covering the Baroque, Classical, and Romantic eras), and instrumentation [6].

The following section provides a collection of representative studies centred on the structural analysis of music. It begins with an overview of research focused on low-level section boundary detection, and then transitions to studies that introduce methodologies for automatically annotating sections and assigning overall form labels (one study).

Many different approaches have been applied to section boundary detection. One of the earliest examples of an unsupervised segmentation algorithm was produced by Mauch et al. [33]. The authors made use of a greedy algorithm to search the SSMs for diagonal path structures that represent repeated sections. Another early approach is the work presented by Turnbull et al. [34] which made use of boosted decision stumps.

Jiang and Müller [35] introduced an automated method for the analysis of specifically first movements of Beethoven piano sonatas (it was known at the start that these movements are in sonata form). An initial coarse analysis, based on audio thumbnailing, was carried out to identify the corner stone sections (exposition and recapitulation), followed by a rule-based method for the finer segmentation.

A common approach is to use a Gaussian checkerboard kernel matrix in conjunction with a self-similarity matrix (SSM) to produce a novelty function over the entire piece [36], [37], [38]. The final segment boundaries of the piece can then be determined by a peak picking algorithm. Clustering methods have also been used for the purpose of boundary detection. Kaiser et al. [39] made use of hierarchical clustering for the task of audio segmentation. Cannam et al. [40] used clustering methods based on timbre-type histograms, where clusters represented segments. Salamon et al. [41] and Wang et al. [42] made use of a technique called spectral clustering for their experiments. Other approaches include the use of linear discriminant analysis (LDA) [43] and multi-resolution community detection [44].

Recently, neural networks (NNs) have become the state-of-the-art approach for automatic boundary segmentation [45]. More specifically, CNNs [46] have been applied to perform this task. Generally, this is achieved by windowing the audio and calculating two-dimensional features that are then passed on to a CNN. The two-dimensional features are either audio features or SSMs and self-similarity lag matrices (SSLMs). These networks generally have only one output node which represents the probability that the provided window contains a section boundary [45], [47], [48], [49], [50]. Common features include: chromagrams [33], [36], [37], mel-scaled log-magnitude spectrograms (MLSs) [42], [45], [47], [48], [49], [50], constant-Q transform (CQT) spectrograms [41], [43], [44] and mel-frequency cepstrum coefficients (MFCCs) [36], [37], [41], [44]. These features are then often used to calculate SSMs [33], [36], [38], [39], [44] and SSLMs [45], [49], [50].

Allegraud et al. [51] made use of a hidden Markov model (HMM) for the purpose of learning musical structure, specifically to learn the structure of 32 sonata form movements from string quartets composed by Mozart. Distinctive sections of the sonata form were encoded as hidden states in the HMM at various levels of detail. By combining manual annotations of the sonata sections with binary analysis features (that indicate the presence of musical properties), the parameters of the HMMs were calculated. The final model was able to compute section boundaries within the pieces as well as determine the type of each identified segment.

Efforts have been devoted to applying computational methodologies towards the enhancement of human musicological analyses. Gotham and Ireland [52] proposed standards (amenable to computational processing) for the representation/visualization of musical form obtained through human analyses. Weiss et al. [53] used automatic visualization of diatonic content unfolding over time in audio recordings to aid human musicological fine-grained structural analyses of the first movements (in sonata form) of early Beethoven piano sonatas. Relatedly, a new dataset dedicated to the first movements of all of Beethoven's piano sonatas was introduced by Zeitler et al. [54], containing *inter alia* symbolic scores, audio data, alignments, and chord and local key information.

As to directly classifying the form of a musical piece, Szelogowski et al. [6] were the first to propose a methodology. They used the SMFSA [32] dataset along with TreeGrad, a method that applies gradient-boosted decision trees as neural networks to achieve their results. The classifier took as input column-wise averaged SSMs, with the duration information appended. In addition, the authors conducted phrase analysis on their dataset, using a peak-picking algorithm to segment the music. A bidirectional long short-term memory (LSTM) network was then trained to annotate the segments.

IV Dataset

In this study, we make use of a custom dataset that was developed for the purpose of overall form recognition. The dataset contains form labels for piano sonatas composed during the Classical period. This period in music occurred roughly between 1730 and 1800 according to [55], while [56] places the start at roughly 1759 (Haydn's first symphony) and the end at 1828 (the death of Schubert). It was a period marked by formal discipline, in contrast to what might be deemed "romantic" [57]. Pieces from this period generally contain the characteristics of order, hierarchy and unambiguous boundaries between different sections. These factors make the Classical period particularly suitable for studying automatic form recognition, as the structure of pieces from this period is generally more clearly defined.

The dataset comprises sonata movements by prominent composers of the period, including Mozart (1756-1791), Clementi (1752-1832), Beethoven (1770-1827), Czerny (1791-1857), and Haydn (1732-1809). Works by composers like Czerny may be classified as late-Classical.[1] Each movement in the dataset contains one or more form labels from musicological sources. The most common form labels from the period, as revealed by the dataset, were discussed in Section II. By using form labels from analyses performed in the published literature, the validity of the labels is ensured. Books (monographs) [7], [8], [9], [12], [13], [16], [17], [18], [19] provided the form labels for pieces composed by Mozart, Beethoven and Haydn, whereas theses [20], [21] provided the labels for the compositions by Clementi and Czerny. Table I gives the label sources for each composer included in this study. Initially, the dataset contained 501 movements to which 821 labels were assigned. After we removed movements that did not have suitable audio and labels for which there were not enough examples, the dataset had 480 movements with 784 associated form labels, and was used to produce the main results of Tables III to V. Appendix A shows the distribution of labels for composers and form types.[2] A notable number of movements were also given different form labels by different musicologists (12% of the dataset).

TABLE I Musicological Sources Used for Sourcing Form Labels for Piano Sonatas by Composers From the Classical Period

Composer	Sources
Mozart	E. Stainkamph [9], J. Salsbury [7], F.H. Marks [8]
Beethoven	D.F. Tovey [12], E. Stainkamph [13], H.A. Harding [17]
Haydn	E. Stainkamph [16], S. Foster [18], B.A. Brown [19]
Clementi	T.E.K. Radloff [20]
Czerny	L. Larson [21]

TABLE II Table Showing the Performers of the Piano Sonata Recordings Used in the Experiments

Composer	Performer
Mozart	Mitsuko Uchida
Beethoven	Daniel Barenboim
Haydn	Rudolf Buchbinder
Clementi	Costantino Mastroprimiano
Czerny	Martin Jones

TABLE III Overall Performance Comparison of the Different Input Features Used to Fit Neural Network Models

Feature	Macro		Weighted		Ranking Loss	Coverage
	AUC	AP	AUC	AP		
Velocity piano roll SSM	**0.823**	0.454	**0.858**	**0.646**	**0.125**	**2.045**
Binary piano roll SSM	0.797	0.448	0.838	0.629	0.129	2.070
T.I. velocity piano roll SSM[a]	0.809	**0.470**	0.843	0.638	0.127	2.066
T.I. velocity piano roll SSLM[a]	0.796	0.432	0.844	0.623	**0.125**	2.047
Velocity piano roll SSLM	0.791	0.446	0.827	0.631	0.132	2.099
Melody piano roll SSM	0.788	0.404	0.827	0.603	0.153	2.239
Mel-spectrogram SSM	0.774	0.396	0.818	0.600	0.154	2.274
Chroma SSM	0.775	0.391	0.803	0.576	0.156	2.268
Tonnetz SSM	0.779	0.399	0.806	0.580	0.163	2.336
Cyclic tempogram SSM	0.483	0.145	0.476	0.275	0.224	2.738
Combination SSMs[b]	0.811	**0.470**	0.841	0.642	0.138	2.142
Szelogowski et al. vector [6]	0.587	0.192	0.648	0.387	0.210	2.655

[a] T.I. refers to transposition invariant

[b] Combination refers to a multi-channel input including velocity piano rolls, cyclic tempograms, Tonnetz and transposition invariant SSMs

All model outputs were generated according to the stratified 20-fold cross-validation methodology and metrics for the outputs calculated. The metrics indicate the overall performance of the form recognition system for different input features. Optimal values are given in bold.

The audio recordings used for this study were sourced from YouTube and Spotify. All features used in this study were derived from the audio recordings through automatic transcription and signal processing, as described below. Since no original symbolic scores accompany the audio recordings, piano roll features were automatically extracted via audio transcription. Table II provides the performers whose recordings were used in this study for each composer (one recording per movement was used).

V Methodology

To identify the form of a musical piece, a concise feature representation is needed that is distinctive to particular forms. As shown in the literature review, the combination of CNNs and windowed SSMs is currently state-of-the-art when it comes to boundary detection and forms the core of many segmentation approaches. Due to the fact that CNNs seem able to produce segment boundaries with a fair degree of accuracy, it is reasonable to anticipate that the identified segment boundaries could be used in subsequent layers of the network to determine the overall form of the piece. SSMs can be generated from a multitude of different features. Feature extraction is used to convert the audio to a new variable space in which the problem is

hopefully easier to solve and is also used to speed up computation in cases where the original input features are computationally infeasible [58]. The goal is to reduce the dimensionality of the input that is given to the machine learning algorithm while retaining the discriminatory information. In the case of machine learning algorithms that are applied to the domain of music, features will generally highlight different aspects of the music being processed.

A. Piano Roll Features

We made use of piano roll features, which appears to be far less common in the literature surrounding this topic. Piano rolls are represented by a two-dimensional matrix where one dimension consists of the 88 notes of a standard piano and the other dimension is discrete time. A piano roll is a representation that indicates the notes that should be sounding at every time instant of the piece of music. The piano rolls were calculated by making use of a neural network developed by Kong et al. [59], which has been state-of-the-art at the task of transcribing audio of piano performances to the musical instrument digital interface (MIDI) format.[3] Their approach achieved a note onset F1-score (the harmonic mean of precision and recall) of 96.72% and a pedal onset F1-score of 91.86% on the MAESTRO dataset [60]. To generate MIDI files from the audio, we employed this model with its default parameters: 768 ticks per second; 160000 segment samples; 100 frames per second; and onset, offset, frame, and pedal offset thresholds of 0.3, 0.3, 0.1, and 0.2 respectively. Note onset, offset, and velocity values in the MIDI data were used to generate different types of piano roll features. Velocity is a measure of how much force the key is pressed with (in the case of keyboard instruments) which relates to the perceived loudness. MIDI derived piano rolls reflect choices made by the performers with respect to expressive timing and dynamics laid out in the original music score. We generated three types of piano rolls from the MIDI data. First, we created binary piano rolls that indicate whether a note is being played or not. Next, we developed velocity piano rolls, which extend the binary version by including the velocity values from the "note on" MIDI messages. Finally, melody piano rolls were produced to approximate the removal of accompanying chords by retaining only the velocity values of the highest notes played in each column, discarding all others. The rationale for this approach is that the melody is often contained in the highest-pitched notes being played, with the accompanying chords occupying the lower pitches. Isolating the melody may facilitate identifying structural elements such as motif development, repetition, variation, or contrast. These are key indicators when analysing musical form.

B. Audio Features

In addition to the above piano roll features, we also investigated the use of common audio features, namely chroma features, mel-spectrograms, Tonnetz features, and cyclic tempograms; most of these features have previously been used with SSMs in the context of music structure analysis, e.g., [11]. These audio features were extracted using librosa [61].

To produce a mel-spectrogram a power spectrum is calculated from the discrete Fourier transform (DFT) coefficients for each window and passed through a set of filter banks spaced in accordance with the mel scale. The mel scale is used as it has been shown to more accurately represent how humans perceive and interpret pitch [62]. The mel-spectrograms were computed at a sample rate of 22050 Hz, with a window size of 2048 samples, a hop length of 512 samples, a Hann window function, power scaling set to 2, and 128 mel bins.

Chromagram features aim to condense the spectral information given by a spectrogram into a pitch representation consisting of 12 pitch classes, one for each distinct note of the Western musical scale. The chroma features were computed at a sample rate of 22050 Hz, with a window size of 2048 samples, a hop length of 512 samples, 12 chroma bins, a Hann window function, and normalized using the maximum norm.

Tonnetz features offer a compact representation of the harmonic content of a signal by projecting its chroma features into 6-dimensional space. Tonnetz (from tone network) is a grid diagram that represents the tonal relationships between pitches. When octave and enharmonic equivalence are assumed, the planar Tonnetz map becomes a hypertorus [63]. Harte et al. [63] developed a method whereby the 12 notes present in a chromagram are projected onto a six-dimensional interior space contained by the surface of the hypertorus. The perfect fifth, major third and minor third are each afforded two coordinates in this representation. The mapping is given by

$$\zeta_n(d) = \frac{1}{||c_n||_1} \sum_{l=0}^{11} \Phi(d,l) c_n(l), \quad \begin{cases} 0 \leq d \leq 5 \\ [0.3em] 0 \leq l \leq 11 \end{cases}, \tag{1}$$

where is c_n is the nth chroma vector, l is the chroma vector pitch class index, and d is an index representing one of the six dimensions of $\zeta_n(d)$ [63]. Φ is the transformation matrix; see [63, Eq. (2)]. The six-dimensional tonal space can be represented as three circles. There is a circle for the perfect fifth, minor third and major third; following Harte et al. [63], the radii were set to 1, 1 and 0.5 respectively.

Cyclic tempograms are robust mid-level tempo representations that capture tempo information of the music being analysed. In this case, the complete tempo spectrum is distilled into a representation that highlights tempi differing by a power of two [64]. The novelty spectrum was computed with a window size of 2048 samples, a hop size of 512 samples, and a gamma parameter of 100 then resampled to 100 Hz. Fourier and autocorrelation-based tempograms were calculated with a window size of 500 samples and a hop size of 50 samples over a tempo range from 30 to 600 beats per minute (BPM). Finally, the cyclic representation was obtained by folding the tempo axis logarithmically using a reference tempo of 30 BPM, 15 bins per tempo octave, and a total of 4 tempo octaves.

C. Downsampling

Because of the excessively large size of the resulting feature matrices, it is necessary to perform size reductions along the time axis of the matrix. The SSMs that would result from the feature vectors without downsampling would not be computationally feasible for a CNN to process. Another requirement of CNNs is that a standard input dimensionality is used. For this reason, the Pillow library [65] for Python was used to resample the feature matrices. The feature matrices were resampled to a time resolution of 100 samples using bilinear interpolation, while the resolution of the feature axis was left unaltered. The use of 100 time samples is a compromise between computational feasibility and SSM detail. This resolution was selected because it still showed sufficient detail when visually inspected for a sample of movements. Feature vectors were then normalised by calculating the Euclidean norm for each time slice of the feature sequence [11]. This is done to encode relative differences rather than absolute differences in the feature representations, which makes the features invariant to differences in sound intensity [11].

D. Self-Similarity Matrices

Once the feature vectors have been produced, the SSMs can be calculated. A SSM captures internal relationships in a series of data by comparing each segment to every other segment in the series [66], resulting in a representation that is fundamental to the analysis of musical structure [11]. Given a feature sequence $X = (x_1, x_2, \ldots, x_N)$, where N is the number of time slices, the SSM $\mathbf{S} \in \mathbf{R}^{N \times N}$ is then given by

$$\mathbf{S}(n,m) := s\left(x_n, x_m\right), \tag{2}$$

where $x_n, x_m \in X$ for $n, m \in [1:N]$, n and m are the row and column indices of $\mathbf{S}$, and s is a function used to calculate the similarity between feature vectors [11]. x_n and x_m are the nth and mth element of X respectively. For this study, we used Euclidean distance as a measure of similarity, which has been shown to perform better than cosine similarity at the task of segmentation [42]. SSMs contain block and path structures. When the feature sequence captures musical characteristics that remain relatively stable throughout a section of music block-like structures occur. Blocks are defined by two segments $\alpha = [s:t]$ and $\alpha' = [s':t']$, where s is the starting index of the segment and t is the ending index of the segment. The block B is the subset

$$B = \alpha' \times \alpha \subseteq [1:N] \times [1:N], \tag{3}$$

where $\subseteq$ denotes a subset and N is the number of rows and the number of columns in the SSM [11]. In contrast, path-like structures arise from recurring subsequences within the feature sequence [11]. The path P of length L is the sequence

$$P = ((n_1, m_1), \ldots, (n_L, m_L)) \tag{4}$$

of cells (n_l, m_l) where $l \in [1 : L]$, with $m_1 = s$ and $m_L = t$ and $(n_{l+1}, m_{l+1}) - (n_l, m_l) \in \Sigma$, where Σ is the set of allowed step sizes [11].

Fig. 1 shows an example of an SSM that was calculated from velocity piano roll features that were downsampled to 100 time instances. Section annotations have been added that indicate where a listener would place the start of each section along with a letter that indicates which sections are new and which are being repeated. Mozart's Rondo Alla Turca (K. 331, 3 rd movement) is slightly different from the standard form of a rondo. The standard form starts with a section denoted by A which is repeated three times as a refrain: ABACA. As can be seen from Fig. 1, the form of the Rondo Alla Turca is given by ABCBABD. It once again contains a refrain B which is repeated three times but rather than starting with the refrain, as in the standard form, it starts with an episode A which is followed by the refrain B. A further difference is that the Rondo Alla Turca is finished by a coda D (tail section).

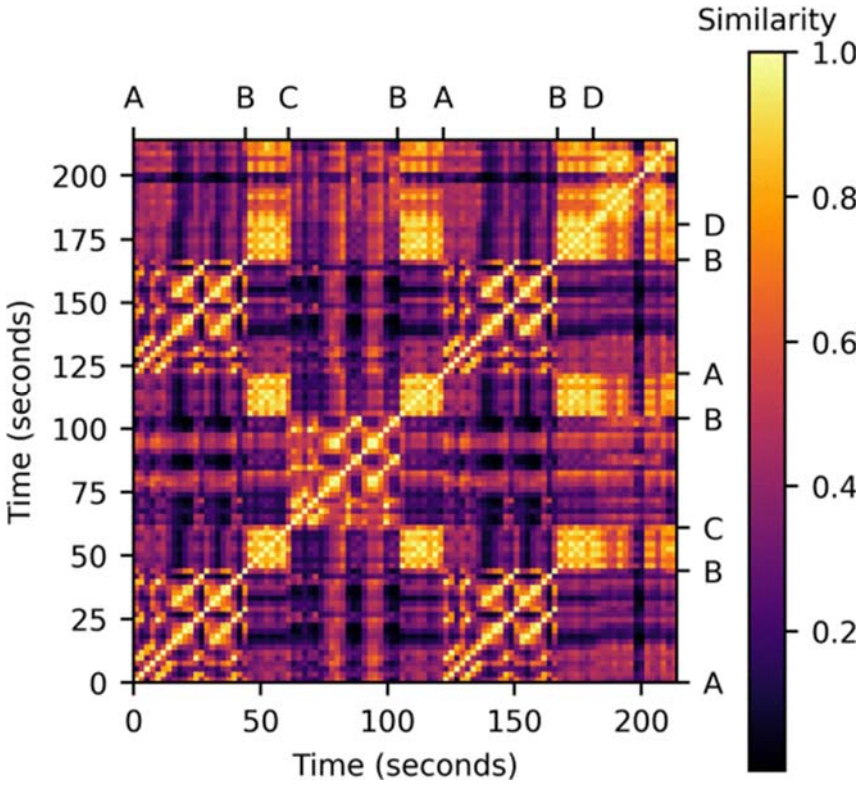

Fig. 1.
SSM derived from velocity piano roll data for W.A. Mozart, Sonata in A Minor, K. 331, 3 rd movement (Rondo Alla Turca). A, B, C and D indicate the start of sections in the movement.

Our work exhaustively investigates classifiers that take standard SSMs as input. In an ancillary step, the best performing SSM features were used to calculate SSLMs that were likewise used as classifier inputs, for purposes of comparison. Unlike a standard SSM, which compares all pairs of absolute time points, a self-similarity lag matrix (SSLM) organises similarity by temporal lag, emphasising how features recur over specific time intervals. Assume that time frames are indexed starting at $n = 0$, and that $X = (x_0, x_1, \ldots, x_{N-1})$ and the SSM is indexed by $[0 : N - 1] \times [0 : N - 1]$. Then the time-lag representation $\mathbf{L}$ of $\mathbf{S}$ is given by

$$\mathbf{L}(\ell, n) = \mathbf{S}(n + \ell, n) \tag{5}$$

for $n \in [0 : N - 1]$ and $\ell \in [-n : N - 1 - n]$ [11]. In practice, the circular form of the SSLM is often used, as it retains the dimensions of the SSM; see [11, Eq. (4.45)].

E. Model Architecture and Training Parameters

The above two-dimensional matrices (SSMs and SSLMs with a dimensionality of 100 by 100) were fed directly into a CNN. The CNN consisted of three convolutional layers. After each convolutional layer there was a max-pooling layer. The discrete convolution *O* of the two-dimensional image *I* and the two-dimensional filter *K* is defined by

$$O(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n), \tag{6}$$

where i and j are indices of O, m is the height of the filter K, and n is the width of the filter K [67]. A convolutional filter is capable of configuring itself in such a way that it becomes an edge detector [68] and produces a feature map of where a transition from one section to another takes place. The max-pooling layers were used for dimensionality reduction. Deeper layers of the CNN can then use the path and block feature maps to infer the form of the provided example. Each of the three convolutional layers contained 16 filters with a width and height of three. The max-pooling filters had a height and width of two. No padding was applied in the network architecture. A diagram of the network architecture can be seen in Fig. 2. The output from the final layer was flattened and passed to a dense layer containing 64 nodes, before concluding with the final classification nodes. Each label was allocated a node of the final output layer. All layers utilised the rectified linear unit (ReLU) activation function, except for the final decision layer, which employed the sigmoid function. The model was trained using the Adam optimiser [69] and the loss function employed was binary cross-entropy L given by

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[y_{ij}\log(p_{ij}) + (1-y_{ij})\log(1-p_{ij})\right], \tag{7}$$

where N represents the number of samples, M is the number of labels, y_{ij} the binary label for the ith target sample and the jth label, and p_{ij} is the predicted probability of the ith sample and the jth label [70]. We observed that the model was prone to overtraining. To prevent this, dropout regularisation (with a rate of 0.25) was used after each convolutional layer and each dense layer except for the network output layer. In addition, L1 and L2 regularisation were used on the CNN layers with values of $1E-5$ and $1E-4$. The model was implemented in Python using the Keras library [71], with the default initial learning rate of 0.001. To initialise the weights of the network the default Keras weight initialiser was used, which is the Glorot uniform initialiser. Additionally, the default Keras bias initialiser, which sets all bias values to zero, was used. An attempt was made to optimise the model architecture and hyperparameters through Bayesian optimisation. However, this method produced results that were inferior to the results obtained by manually selecting architectures and parameters. This outcome can likely be attributed to the limited exploration of the model design space, constrained by the substantial evaluation time required for each architecture. A process of one-hot encoding was used to convert the set of labels to a set of target vectors for the model.

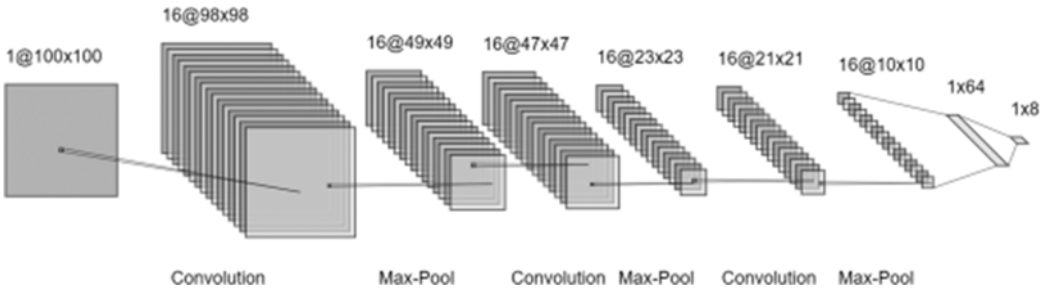

Fig. 2.
The CNN architecture used for this study, showing the dimensionality of the outputs of each convolutional and max-pooling operation. Each convolutional layer makes use of 16 kernels with a width and height of three; max-pooling uses two by two kernels. There is a reduction in the size of the feature maps after each layer.

F. Comparison With Szelogowski et al.

Szelogowski et al. [6] is the only other study known to the authors that focused on overall form recognition. Their methodology could not be replicated exactly as the TreeGrad algorithm that was used is not able to output multiple labels. Instead, a standard artificial neural network (ANN) was used to test the features. The model architecture was selected to ensure that the number of trainable parameters remained roughly equal between the fixed CNN model and the ANN model. This resulted in an ANN that consisted of four fully connected layers, consisting of 40, 32, 16 and 8 neurons respectively. Dropout regularisation (with a rate of 0.25) was included after all of the layers except the final layer. ReLU activation functions were used on all of the neurons except the final layer, which made use of the sigmoid activation function. All of the other model parameters and training methods were kept constant.

G. Testing Procedure

Due to the relatively small dataset, the experiments were performed by stratified multi-label *k*-fold cross-validation [72]. Stratified *k*-fold cross-validation is a technique whereby a dataset is subdivided in a way that retains the overall class/label distribution in each of the subdivisions. In the *k*-fold testing methodology, each subdivision/fold is used in turn as a test set, while the excluded data is used as training data. This testing methodology has the benefit of allowing each sample in the dataset to act as a test sample. Twenty folds were used in an attempt to maximise the aggregate amount of training data used in each test fold. From the training data, 15% was taken for a validation set. The validation set guided the training by providing a metric whereby early stopping could be implemented. With a patience of 100 epochs, the model was trained for 1000 epochs. Early stopping was based on monitoring the loss value of the validation set. If no improvement in the validation loss was observed for 100 epochs the training would stop and the model weights would be restored to the values that yielded the lowest loss on the validation set. A batch size of 64 was used during training. The final trained model was then tested on the hold-out/test set as per the cross-validation method.

VI. Results

A. Overall Performance Evaluation

Table III presents the overall performance metrics derived from the stratified 20-fold cross-validation testing process. A mathematical description of these multi-label metrics can be found in [73]. In short, the area under the curve (AUC) is computed by finding the area under the curve formed by plotting the true positive rate (TPR) and false positive rate (FPR) at various threshold values. The average precision (AP) is determined by calculating the area under the curve formed by plotting precision and recall at different threshold values. Ranking loss indicates the frequency with which irrelevant labels are ranked higher than relevant ones. The coverage metric measures how far, on average, one must go down the ranked list to include all relevant labels for an example.

We considered both macro and weighted averages for this study. Macro averaging computes the metric independently for each class and then takes the unweighted mean across all classes, whereas weighted averaging computes the metric for each class and takes the mean weighted by the number of true instances (support) in each class. Macro averages are more important for this application as all label scores contribute equally to the final performance measure, which is not the case in a weighted average where class imbalances can dramatically alter the results. These results indicate small differences between velocity piano rolls, binary piano rolls and the multi-channel input as the three inputs score comparably over multiple metrics. For the velocity and binary piano rolls the macro-AUC indicates good predictive ability. In almost all cases, the weighted-AUC is higher as it gives more weight to the sonata labels, which the model is better at determining (according to Table IV). The macro-AP metric demonstrates that SSMs calculated from transposition-invariant velocity piano rolls outperform standard velocity piano rolls. Velocity piano rolls generate the highest quality ranked lists of labels, according to the ranking loss and coverage metrics. The coverage value indicates that on average only the first two labels from the ranked classifier output are needed to fully cover the target labels for each movement. Table III indicates that the SSMs based on melody piano rolls did not perform to the same degree as those based on velocity piano rolls. This suggests that the accompanying chords are also important

for determining the structure of a movement, as anticipated. SSMs based on the Tonnetz, chromagram, and mel-spectrogram did not perform as effectively as those based on piano rolls. This suggests that converting movements into a detailed note-based representation, like a piano roll, is beneficial for identifying musical form. Table III indicates that cyclic tempogram-based SSMs lack significant predictive power. In an ancillary experiment (noted earlier), SSLMs were computed for the best-performing SSM features, namely velocity piano rolls. The SSLMs performed worse than their SSM counterparts for all metrics, except for ranking loss, where a tie was observed between velocity piano roll SSMs and transposition invariant velocity piano roll SSLMs.

TABLE IV Comparison of AUC Scores Achieved for Each Label When Using Features Listed in Table III

Feature	Binary	Sonata	Ternary	Rondo	Minuet and Trio	Rondo-Sonata	Scherzo and Trio	Theme and Variations
Velocity piano roll SSM	0.636	0.903	**0.839**	0.848	0.822	**0.896**	0.818	0.825
Binary piano roll SSM	0.477	0.902	0.790	0.813	0.841	0.895	0.820	0.833
T.I. velocity piano roll SSM[a]	0.553	0.897	0.774	**0.861**	**0.845**	0.794	**0.906**	0.844
T.I. velocity piano roll SSLM[a]	0.597	**0.913**	0.792	0.854	0.814	0.714	0.810	**0.873**
Velocity piano roll SSLM	0.486	0.906	0.822	0.808	0.829	0.861	0.816	0.800
Melody piano roll SSM	0.528	0.883	0.823	0.780	0.832	0.817	0.820	0.818
Mel-spectrogram SSM	**0.714**	0.883	0.825	0.748	0.818	0.656	0.800	0.747
Chroma SSM	0.6864	0.848	0.8054	0.744	0.8215	0.7242	0.8033	0.7737
Tonnetz SSM	0.654	0.850	0.788	0.782	0.774	0.805	0.733	0.850
Cyclic tempogram SSM	0.503	0.462	0.503	0.464	0.453	0.515	0.516	0.451
Combination SSMs[b]	0.661	0.891	0.821	0.806	0.810	0.834	0.815	0.8454
Szelogowski et al. vector [6]	0.505	0.713	0.656	0.606	0.711	0.468	0.674	0.365

[a] T.I. refers to transposition invariant.

[b] Combination refers to a multi-channel input including velocity piano rolls, cyclic tempograms, Tonnetz and transposition invariant velocity piano roll SSMs.

The maximum AUC score for each label is given in bold.

B. Performance With Respect to Individual Labels

Table IV shows the AUC scores for each label across different SSM types. As these are individual label scores no averaging is required. Table IV highlights the fact that the model has the most difficulty with the binary form type. This may be due to it having the second fewest samples in the dataset. Interestingly, the mel-spectrograms perform best for the binary label, although with such a small sample size this could be a statistical anomaly.

The sonata label column generally contains higher scores when compared to other labels for a given SSM type. This is to be expected because sonata labels make up the majority of the label dataset. For movements in sonata form the exposition is often repeated in the score and the performance. This creates a distinctive path structure running parallel to the main diagonal in the first section of the SSM, which could aid the model in identifying the sonata form in movements. Interestingly, the transposition invariant velocity piano roll SSLM obtained the highest score for sonata form, in spite of both the macro and weighted AUC being less than that of the velocity piano roll SSMs (cf. Table III).

All other labels received an AUC score of over 0.8 for the top performing self-similarity matrices. It is encouraging to see relatively high piano roll scores for the theme and variations form type as it also has limited representation in the dataset. This is most likely due to the theme and variations form SSMs being uniquely distinctive in comparison to SSMs based on other form types. Theme and variations form is characterised by stating an unembellished theme and then repeating the theme several times with melodic, rhythmic, harmonic, or other changes.

From Table IV it may appear as if transposition invariant velocity piano roll SSMs are superior to the velocity piano roll SSMs. Transposition invariant velocity piano roll SSMs score the highest individual label scores for the rondo, minuet and trio and scherzo and trio labels. This is contrary to the results in Table III where velocity

piano roll SSMs outperform transposition invariant velocity piano roll SSMs. The discrepancy is explained by noticing that the results of transposition invariant velocity piano roll SSMs are much more varied.

C. Significance Results for SSMs Based on Spectrograms and Piano Rolls

According to the above results, it would appear as if note data in the form of piano rolls is a superior input into the form recognition system compared to features derived from spectrograms. However, sources of randomness such as dropout regularisation and weight initialisation may account for the observed differences in the metrics. Experiments were performed to determine if the differences in the results that were observed between SSMs based on velocity piano rolls and those based on mel-spectrogram SSMs were statistical artifacts. Inspired by Velarde et al. [74], the Wilcoxon signed-rank test [75] was used to do this.

The null hypothesis is that there is no difference in performance between the two different SSM input feature types. The alternative hypothesis is that velocity piano roll-based SSMs perform better than mel-spectrogram SSMs at each of the calculated metrics. A one-tailed test was used, as earlier results already indicated superior performance of velocity piano roll SSMs.

To perform the Wilcoxon signed-rank test, samples are required from the two populations that need to be compared. A similar training protocol to the previously defined protocol was used to generate the samples, with the main difference being how the data was split. A simple randomised train-test split was used with 15 percent of the total data being used for a test set and 15 percent of the training data being used as a validation set. The same model architecture, parameters and early stopping criteria were used. To accumulate the required samples 100 randomised train-test splits were generated.

For each split, one model was trained and tested on velocity piano roll SSMs and the other on mel-spectrogram SSMs. Once training was completed, the model was used to make predictions on the test set. The predictions and ground truth were then used to calculate the metrics in Table IV for each model for the given split. This resulted in 100 paired samples of each test statistic for the two different types of SSMs. From these samples, the Wilcoxon test statistic and associated *p*-value were calculated. Table V shows the one-tailed *p*-values for the different metrics that resulted from performing the Wilcoxon test on the generated samples. The *p*-values can be interpreted as the probability of observing the collected metrics if the null hypothesis were true.

TABLE V The *p*-Values for a One-Sided Wilcoxon Signed-Rank Test and Adjusted *t*-Test Comparing the Performance of Velocity Piano Roll SSMs and Mel-Spectrogram SSMs as Inputs for the Form Recognition System

Metric	Wilcoxon	Adjusted t-Test
Macro-AUC	4.03×10^{-12}	0.038
Weighted-AUC	2.87×10^{-11}	0.044
Macro-AP	1.69×10^{-13}	0.024
Weighted-AP	2.07×10^{-10}	0.053
Ranking Loss	5.82×10^{-14}	0.018
Coverage	1.29×10^{-14}	0.014

Results are based on 100 simulations for each input type.

The Wilcoxon test requires independently sampled data, which was not strictly satisfied due to repeated resampling. To address this, we also report an adjusted *t*-test following Nadeau and Bengio [76], who proposed an adjusted version of the Student's *t*-test [77], which is claimed to be less biased when the requirement of independent samples is violated. Table V shows the one-tailed *p*-values for the different metrics.

Table V shows that all Wilcoxon test *p*-values indicate statistical significance at the $p<0.05$ level. Adjusted *t*-tests are also significant except for weighted-AP, which slightly exceeds the threshold. Therefore, the null hypothesis can be rejected in favour of the alternative, demonstrating that velocity piano roll SSMs outperform mel-spectrogram SSMs.

D. Influence of Performer on Model Outputs

Pianists strive to create a uniquely personalised interpretation of a musical score. The differences in interpretation are often subtle and nuanced. To ensure that the model was not biased towards certain pianists or stylistic choices, additional experiments were conducted to assess its sensitivity to performance variations. The test set contained five different performances of the three movements of Beethoven's Sonata No. 8, also known as the "Pathétique". Certain Barenboim movement recordings, previously used in other experiments, were included and supplemented with recordings by other pianists. The entire dataset (excluding movements in the above test set) was used to train the model, with velocity piano roll SSMs serving as input features. It is worth emphasising that the movements of Beethoven's Sonata No. 8 were not in the training set. The model was therefore inferring the form of a piece that it had never "seen" before.

Velocity piano rolls were selected as this input feature performed well in several metrics (see Table III). In addition, velocity piano rolls provide a better representation of a performer's interpretation than a binary piano roll. Binary piano rolls would account for expressive timing choices made by the performer but not for dynamics. The different performances were then input into the model and the output neuron activations were recorded. Neurons in the output layer generate continuous values between zero and one. Neuron activations close to one indicate a high likelihood of the label being relevant to the piece, while a value near zero indicates a low likelihood. Table VI provides model outputs for a number of performances of different movements of Beethoven's Sonata No. 8. Ideally, the model should output similar activations for different performances of the same piece of music. This would indicate that the model is performer invariant. Performance durations are also provided to confirm the uniqueness of each performance.

TABLE VI Output Neuron Activations of a Model That Uses Piano Roll SSM Features When Presented With Different Performers' Versions of the Movements of Beethoven's Sonata No. 8

Movement and (Labels)	Performer	Duration	Binary	Sonata	Ternary	Rondo	Minuet and Trio	Rondo-Sonata	Scherzo and Trio	Theme and Variations
Piano Sonata No. 8 in C Minor, Op. 13 - I. (Sonata)	D. Barenboim	9:29	0.041	**0.719**	0.107	0.396	0.093	0.038	0.096	0.044
	F. Say	8:17	0.052	**0.765**	0.078	0.283	0.057	0.023	0.082	0.025
	J. Jandó	8:29	0.026	**0.691**	0.127	0.448	0.096	0.036	0.078	0.061
	S. Hanai	8:29	0.033	**0.670**	0.119	0.499	0.081	0.044	0.094	0.063
	S. Kovacevich	8:47	0.028	**0.674**	0.127	0.457	0.098	0.036	0.090	0.062
Piano Sonata No. 8 in C Minor, Op. 13 - II. (Rondo)	D. Barenboim	5:08	0.127	0.049	0.305	**0.651**	0.156	0.155	0.088	0.038
	F. Say	5:25	0.119	0.045	0.376	**0.607**	0.180	0.154	0.108	0.044
	J. Jandó	4:42	0.133	0.043	0.384	**0.589**	0.158	0.124	0.154	0.039
	S. Hanai	5:22	0.153	0.039	**0.458**	0.418	0.171	0.082	0.226	0.060
	S. Kovacevich	5:33	0.131	0.043	0.429	**0.537**	0.145	0.139	0.150	0.051
Piano Sonata No. 8 in C Minor, Op. 13 - III. (Rondo Sonata, Rondo)	D. Barenboim	4:58	0.182	0.400	0.093	**0.678**	0.038	0.390	0.039	0.065
	F. Say	4:36	0.131	0.452	0.074	**0.761**	0.026	0.332	0.036	0.056
	J. Jandó	4:30	0.154	0.361	0.079	**0.726**	0.029	0.401	0.025	0.045
	S. Hanai	4:44	0.181	0.297	0.119	**0.752**	0.051	0.450	0.057	0.098
	S. Kovacevich	4:17	0.143	0.352	0.087	**0.787**	0.033	0.449	0.034	0.077

For each performer, the maximum activation is given in bold

The first movement of the piece is in sonata form, and the model assigns the highest activation to the output neuron corresponding to this form for all five performers. Fairly high activations are also assigned to the output neuron that corresponds to the rondo form. It is reassuring to note that the outputs for binary, ternary, minuet and trio, rondo sonata, scherzo and trio and theme and variations forms are all low.

The second movement is in rondo form according to musicological sources [12], [13], [17]. In four of the five examples, the highest output for the model corresponds to the rondo form. The model outputs the highest activation for ternary form for S. Hanai's performance. On further inspection, it can be observed that all of the performances of the second movement have fairly high activations for ternary form. Other model form activations are fairly low by comparison.

The third and final movement has been described as having a rondo sonata form by one musicologist [13] and a rondo form by another [12]. In all five instances, the model assigns the highest activation to the rondo label. Fairly high activations can also be observed for the sonata and rondo sonata outputs, giving credence to the idea that the piece can be interpreted as having either rondo sonata or rondo form. Other model outputs are low by comparison, which further indicates that the model is functioning correctly.

E. Influence of Adherence to Repeats on Model Outputs

A much more significant difference between performances is where the pianist deviates from the score. Some performers adhere to the repeats notated in the score, while others opt not to play the repeats or elect to play some repeats and not to play others. For a description of repeats, see Section II. Unique test sets were created to determine the effects on the model outputs of performers adhering to or omitting repeats in their performances.

Beethoven's and Mozart's movements were identified that contained repeats. Three performances of each identified movement were obtained and assessed by a musically trained listener to determine if the repeats were adhered to or not. Choosing movements from the same piece with durations that differed significantly as played by different performers proved to be an effective strategy for finding movements where repeats were performed (or not). A model was trained on the entire original dataset, excluding the movements from the current test set. Once more velocity piano roll SSMs were used as input features. The same model architecture, early stopping methods, and parameters that were described in Section V were used. Different test sets were then input into the model and the output neuron activations were recorded. This experiment was performed twice, once for the Beethoven movements and then subsequently also for the Mozart movements. Tables VII and VIII present the model outputs for each performance of the movements. Ideally, the model would output the same activations for pieces that either adhered to or omitted the repeats notated in the score. This would indicate that the model is repeat invariant. Performance durations and indications as to whether repeats are adhered to are provided to illustrate the uniqueness of each performance.

TABLE VII Output Neuron Activations of a Model That Uses Piano Roll SSM Features When Presented With Different Performers' Versions of Movements (all Having Sonata Form) by Beethoven

Piano Sonata Movement	Performer	Duration	Repeats[a]	Binary	Sonata	Ternary	Rondo	Minuet & trio	Rondo-Sonata	Scherzo & trio	Theme & Variations
No. 15 in D Major, Op. 28 - I.	D. Barenboim	11:25	Yes	0.122	**0.853**	0.082	0.178	0.026	0.032	0.081	0.041
	W. Backhaus	6:42	No	0.106	0.484	0.286	0.342	0.104	**0.497**	0.033	0.026
	W. Kempff	7:17	No	0.133	0.404	0.307	0.315	0.103	**0.431**	0.042	0.034
No. 25 in G Major, Op. 79 - I.	D. Barenboim	3:26	Yes, No	0.058	**0.842**	0.168	0.016	0.004	0.004	0.015	0.005
	R. Brautigam	4:38	Yes, Yes	0.186	**0.504**	0.171	0.349	0.168	0.293	0.035	0.064
	R. Buchbinder	4:34	Yes, Yes	0.145	**0.424**	0.214	0.393	0.133	0.176	0.034	0.032
No. 3 in C Major, Op. 2 - I.	D. Barenboim	10:23	Yes	0.075	**0.831**	0.135	0.110	0.054	0.115	0.028	0.059
	I. Levit	9:40	Yes	0.080	**0.827**	0.144	0.109	0.060	0.159	0.027	0.058
	W. Backhaus	7:22	No, No	0.268	**0.501**	0.304	0.392	0.264	0.391	0.239	0.298
No. 6 In F Major, Op. 10 (2) - I.	D. Barenboim	6:06	Yes ,No	0.077	**0.929**	0.035	0.090	0.021	0.013	0.052	0.023
	I. Levit	5:16	Yes, No	0.110	**0.915**	0.049	0.101	0.022	0.014	0.058	0.015
	P. Takács	9:35	Yes, Yes	0.153	**0.658**	0.248	0.103	0.159	0.161	0.055	0.049
No. 9 in E Major, Op. 14 - I.	D. Barenboim	6:49	Yes	0.040	**0.974**	0.034	0.035	0.008	0.063	0.002	0.001
	I. Levit	6:13	Yes	0.028	**0.980**	0.028	0.028	0.005	0.044	0.002	0.001
	W. Backhaus	5:12	No	0.161	**0.744**	0.206	0.304	0.111	0.216	0.048	0.029

[a] This column indicates adherence to notated repeats within the movement by the performer. Each repeat is assigned either a "Yes" or a "No" in order of appearance depending on whether it was adhered to or not. For example, if a movement was designated "Yes,No" it means that there are two notated repeats in the score and that the first repeat was played and the second repeat was not played during the performance.
For each performer, the maximum activation is given in bold.

TABLE VIII Output Neuron Activations of a Model That Uses Piano Roll SSM Features When Presented With Different Performers' Versions of Movements (all Having Sonata Form) by Mozart

Piano Sonata Movement	Performer	Duration	Repeats[a]	Binary	Sonata	Ternary	Rondo	Minuet & trio	Rondo-Sonata	Scherzo & trio	Theme & Variations
No. 1 in C Major, K. 279 - I	A. Schiff	5:16	Yes	0.007	**0.951**	0.019	0.032	0.006	0.037	0.004	0.014
	G. Gould	4:19	No	0.219	**0.477**	0.190	0.432	0.223	0.302	0.123	0.226
	M.J. Pires	6:59	Yes	0.125	**0.484**	0.092	0.193	0.129	0.160	0.072	0.267
No. 13 In B-Flat Major, K. 333 - I	A. Schiff	7:19	Yes, No	0.064	**0.897**	0.052	0.065	0.022	0.055	0.020	0.055
	J. Jandó	7:22	Yes, No	0.039	**0.926**	0.040	0.050	0.015	0.042	0.013	0.038
	V. Horowitz	10:41	Yes, Yes	0.155	**0.621**	0.052	0.130	0.072	0.154	0.044	0.305
No. 14 In C Minor, K. 457 - I	A. Schiff	7:30	Yes, Yes	0.113	**0.760**	0.060	0.063	0.061	0.060	0.029	0.126
	J. Jandó	5:27	Yes, No	0.013	**0.960**	0.029	0.020	0.012	0.009	0.005	0.007
	W. Backhaus	3:57	No, No	0.137	**0.503**	0.049	0.428	0.045	0.259	0.038	0.078
No. 17 in B-Flat Major, K. 570 - I	A. Schiff	5:38	Yes, No	0.012	**0.957**	0.038	0.037	0.007	0.013	0.008	0.017
	M.J. Pires	8:45	Yes, Yes	0.048	**0.721**	0.037	0.060	0.068	0.055	0.032	0.144
	M. Uchida	5:50	Yes, No	0.013	**0.943**	0.042	0.053	0.008	0.019	0.013	0.024
No. 2 in F Major, K. 280 - I	A. Schiff	4:28	Yes, Yes	0.043	**0.932**	0.057	0.035	0.023	0.026	0.015	0.026
	G. Gould	3:15	No, No	0.089	0.361	0.326	**0.455**	0.088	0.196	0.037	0.063
	M.J. Pires	6:42	Yes, Yes	0.142	**0.659**	0.087	0.138	0.070	0.107	0.066	0.177

[a] See corresponding note in Table VII.
For each performer, the maximum activation is given in bold.

Table VII shows the model outputs for movements by Beethoven for different performers and Table VIII shows the model outputs for movements by Mozart for different performers. All of the movements in Tables VII and VIII are labelled as being only in the sonata form. The sonata form was selected as it was the most common form label in the dataset. In addition to this, sonata form movements tend to be longer and contain more variability in terms of structure compared to other form types.

Tables VII and VIII show that the model still performs well when performances that are substantially different in terms of duration and repeat adherence are input into the model. For the selection of Beethoven sonatas (see Table VII), the highest model outputs are assigned to the sonata label in nearly all instances. The only exceptions are the performances by Backhaus and Kempff of the first movement of sonata number 15. In these instances, the model activation is higher for the rondo sonata form. For the selection of Mozart sonatas (see Table VIII) the highest model outputs are assigned to the sonata label for all the performances with the performance by Gould of the first movement of the sonata No. 2 being the only exception. The model activation is higher for the rondo form in this example.

To further interpret the model outputs, it is necessary to characterise the training. The scores were studied to find an example of a movement that contained notated repeats for each combination of composer and form type. The movements were then assessed by a musically trained listener to determine if the performances adhered to the repeats. Only a sample of the training set was assessed, as characterising the entire training set would have proven time-consuming and costly. Practically all of the repeats are adhered to in the sample selected from the training set, with the only exceptions presented by the first movement of Mozart's fourth sonata (K. 282) and the first movement of Haydn's sonata in B minor (Hob. XVI, No. 32), where the second notated repeats are not played. Table VII also partially characterises the training data. In Table VII it can be seen that Barenboim also opts to play the first repeat in the sample of sonatas that were investigated.

Another observation that can be made when studying Tables VII and VIII is that the performances where the pianists adhere to the first notated repeat have higher sonata label activations. For example, in the performances of the first movement of sonata number 15 by Backhaus and Kempff (see Table VII) the first (and only) notated repeat in the score is not played. These performances have much lower sonata label activations than the performance by Barenboim, who opts to perform the notated repeat. This can also be seen for the first movement of Beethoven's sonata number 3, where Backhaus's performance does not include the notated repeat and received a lower activation for the sonata label than the performances by Levit and Barenboim.

The same observation can be made for sonata movements in Table VIII, which deals with a selection of Mozart's sonata movements. The only performance for which sonata form does not receive the highest activation is Gould's performance of the first movement of Mozart's Sonata No. 2. Gould's performance is much shorter than Schiff and Pires's performances, who play both of the notated repeats in the score, while

Gould does not play either. Furthermore, Backhaus does not play the first notated repeat in the first movement of Mozart's fourteenth piano sonata (K. 457) and correspondingly the model activation for the sonata label is much lower than in the other two performances where the first repeat is played (see Table VIII).

From these observations, it would appear as if the model outputs higher sonata activations for sonata form performances where the first notated repeat is played. This is to be expected as the exposition is generally repeated when sonatas are performed. A likely explanation for this effect is suggested by the fact that the examples used in the training dataset were performances that generally included repeats (or so it would appear given the partial characterisation of the training data). This effect should be reduced by providing the model with more examples where notated repeats are not adhered to.

VII. Conclusion

We showed that the form of piano sonata movements can be determined to a high degree of accuracy using a combination of SSMs and CNNs and that on average only the first two labels are needed from the ranked list of outputs to fully cover the target set of labels. The value of converting audio to a note-based representation was also demonstrated. By performing a Wilcoxon signed-rank test and adjusted *t*-test it was shown that the superior performance of velocity piano rolls compared to mel-spectrogram-based inputs was real and not the result of a statistical aberration. In addition, our method was shown to be robust to performer interpretative choices, and to their adherence or not to notated repeats. The benefits of using full SSMs as input into neural network based frameworks were demonstrated by comparing the results obtained with this approach with results obtained from a system that uses feature vectors derived from SSMs [6].

It is worthwhile to note that musical form is a complex and high-level feature of musical audio that requires expert knowledge and musical training to identify and even then, there can be differences of opinion on form among experts in the field. An argument can be made that the automatic recognition of form is a more difficult task than other MIR tasks such as genre recognition, which many people can perform to some degree without musical training. The ability to recognise musical form is an important part of MIR that can benefit music recommendation systems.

While the approach we followed in this study sets a new benchmark in the field of automatic form recognition, there are at least two possibilities through which the performance of the system could likely be improved even further. Both of them entails increasing the number of training data. The first possibility is to source additional sonata movements with verified form labels—however this may prove challenging, as demonstrated by the difficulty we had finding musicological sources for a number of Haydn piano sonatas. The second possibility is to include multiple performances of the same movement by different pianists during the training stage. This should make the model more invariant to whether notated repeats are played or not, as the model will be exposed during training to more of the possible ways in which pianists can choose to adhere to them.

Finally, it is worth emphasizing that the methodology described in this article could be extended to (especially) piano music of any genre that adhere to standard forms, for example Romantic-period (nineteenth-century) sonatas, and jazz (as noted earlier). This gives our approach wide applicability, as the piano literature is enormous. Relatedly, our methodology could potentially be used to generate musical structure embeddings across diverse music genres, enabling recommendation systems to perform clustering of music based on structural characteristics. Furthermore, form labeling can be used to confirm and/or broadly guide analyses by music students, musicologists, and performers that use analysis as a tool for aiding their memorization of works to be performed in concert.

Appendix A

Dataset Composition

TABLE IX Counts of Eight Fundamental Sonata Movement Form Label Types Assigned to Movements in the Full

Label	Mozart	Beethoven	Czerny	Clementi	Haydn	Total
Sonata	86	141	16	82	52	377
Rondo-Sonata	26	12	5	0	2	45
Binary	2	0	2	9	15	28
Minuet & Trio	6	13	0	0	25	44
Ternary	8	37	8	38	17	108
Rondo	15	54	7	38	10	124
Scherzo & Trio	0	18	12	0	3	33
Theme & Var.[a]	0	9	2	6	8	25
Total[b]	143	284	52	173	132	784
Movements[c]	51	97	50	173	109	480

[a]Theme and variations
[b]Total number of labels per composer
[c]Total number of movements per composer
The table includes totals by label type (rows), totals by composer (columns), and the total number of movements per composer.

References

[1] P. Cano et al., "ISMIR 2004 audio description contest," Music Technol. Group, Universitat Pompeu Fabra, Tech. Rep. MTG-TR-2006-01, Jan. 2006.
[2] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Found. Trends Inf. Retrieval*, vol. 8, no. 2/3, pp. 127–261, Jan. 2014.
[3] A. Whittall, "Form," in *Grove Music Online*, L. Macy, Ed. Oxford, U.K.: Oxford Univ. Press, Jan. 2001, Accessed: Aug. 9, 2025. [Online]. Available: https://doi.org/10.1093/gmo/9781561592630.article.09981
[4] J. Dunsby and A. Whittall, *Music Analysis in Theory and Practice*. London, U.K.: Faber, 1988.
[5] O. Nieto et al., "Audio-based music structure analysis: Current trends, open challenges, and applications," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 3, no. 1, pp. 246–263, Dec. 2020.
[6] D. Szelogowski, L. Mukherjee, and B. Whitcomb, "A novel dataset and deep learning benchmark for classical music form recognition and analysis," in *Proc. 23rd ISMR Conf.,* Bengaluru, India, Dec. 4–8, 2022, pp. 900–907.
[7] J. Salsbury, *Short and Concise Analysis of Mozart's Twenty-Two Pianoforte Sonatas, With a Description of Some of the Various Forms*. London, U.K.: Weekes, 1917.
[8] F. H. Marks, *The Sonata: Its Form and Meaning as Exemplified in the Piano Sonatas by Mozart, a Descriptive Analysis.* London, U.K.: W. Reeves, 1921.
[9] E. Stainkamph, *The Form & Analysis of Mozart's Pianoforte Sonatas*. Melbourne, VIC, Australia: Allan's Music, 1967.
[10] P. A. D. Burger and J. P. Jacobs, "Direct labelling of form of classicalperiod piano sonata movements from audio recordings," in *Proc. 11th Int. Conf. DLFM*, Stellenbosch, South Africa: ACM, Jun. 2024, pp. 1–5.
[11] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, 2nd ed. Berlin, Germany: Springer, 2015.
[12] D. F. Tovey, *A Companion to Beethoven's Pianoforte Sonatas (Bar-to-Bar Analysis).* London, U.K.: Assoc. Board Roy. Sch. Music, 1951.
[13] E. Stainkamph, *Form & Analysis of the Complete Beethoven Piano Sonatas.* Melbourne, VIC, Australia: Allans Music, 1968.
[14] G. Spring and J. Hutcheson, *Musical Form and Analysis: Time, Pattern, Proportion*. Madison, WI, USA: Brown & Benchmark, 1995.
[15] D. Green, *Form in Tonal Music: An Introduction to Analysis*. New York, NY, USA: Holt, Rinehart and Winston, 1979.
[16] E. Stainkamph, *The Form and Analysis of Twenty-Seven Haydn Pianoforte Sonatas*. Melbourne, VIC, Australia: Allans Music, 1970.
[17] H. Harding and P. Fleury, *Analysis of Form: Beethoven's 32 Piano Sonatas*. Scotts Valley, CA, USA: CreateSpace Independent Publishing Platform, 2014.
[18] S. Foster, "Tonal methods of cyclic unification in Haydn's mature keyboard sonatas," Ph.D. dissertation, Sch. Music, Louisiana State Univ. Agricultural Mech. College, Baton Rouge, LA, USA, 1990.

[19] A. P. Brown, *Joseph Haydn's Keyboard Music : Sources and Style*. Bloomington, IN, USA: Indiana Univ. Press, 1986.
[20] T. E. K. Radloff, "The piano sonatas of Muzio Clementi: An investigation into compositional aspects with special emphasis on developments in form and style," Ph.D. dissertation, Dept. Music Musicology, Rhodes Univ., Makhanda, South Africa, 1987.
[21] L. Larson, "An underestimated master: A critical analysis of Carl Czerny's eleven piano sonatas and his contribution to the genre," Ph.D. dissertation, Sch. Music, Univ. Nebraska, Lincoln, NE, USA, 2015.
[22] W. D. Sutcliffe, "Binary form: Definition," in *Grove Music Online*, L. Macy, Ed. Oxford, U.K.: Oxford Univ. Press, Jan. 2001, Accessed: Aug. 9, 2025. [Online]. Available: https://doi.org/10.1093/omo/9781561592630.013.60000200222
[23] W. D. Sutcliffe and M. Tilmouth, "Ternary form," in *Grove Music Online*, L. Macy, Ed. Oxford, U.K.: Oxford Univ. Press, Jan. 2001, Accessed: Aug. 9, 2025. [Online]. Available: https://doi.org/10.1093/gmo/9781561592630.article.27700
[24] J. Hepokoski and W. Darcy, *Elements of Sonata Theory: Norms, Types, and Deformations in the Late-Eighteenth-Century Sonata*. New York, NY, USA: Oxford Univ. Press, 2006.
[25] J. Webster, "Sonata form," New Grove Dictionary Music Musicians, vol. 23, pp. 687–698, Jan. 2001.
[26] R. Hutchinson, *Music Theory for the 21st-Century Classroom* (Open Textbook Library Series). Tacoma, WA, USA: Univ. Puget Sound, 2017.
[27] M. E. Little, "Minuet," in *Grove Music Online*, L. Macy, Ed. Oxford, U.K.: Oxford Univ. Press, Jan. 2001, Accessed: Aug. 9, 2025. [Online]. Available: https://doi.org/10.1093/gmo/9781561592630.article.18751
[28] T. A. Russell and H. Macdonald, "Scherzo," in *Grove Music Online*, L. Macy, Ed. Oxford, U.K.: Oxford Univ. Press, Jan. 2001, Accessed: Aug. 9, 2025. [Online]. Available: https://doi.org/10.1093/gmo/9781561592630.article.24827
[29] M. S. Cole, "Rondo," in *Grove Music Online*, L. Macy, Ed. Oxford, U.K.: Oxford Univ. Press, Jan. 2001, Accessed: Aug. 9, 2025. [Online]. Available: https://doi.org/10.1093/gmo/9781561592630.article.23787
[30] M. S. Cole, "Sonata-rondo, the formulation of a theoretical concept in the 18th and 19th centuries," *Musical Quart*., vol. 55, no. 2, pp. 180–192, Apr. 1969.
[31] J. Smith, J. Burgoyne, I. Fujinaga, D. De Roure, and J. Downie, "Design and creation of a large-scale database of structural annotations," in *Proc. 12th Int. Soc. Music Inf. Retrieval Conf.,* Miami, FL, USA, Oct. 2011, pp. 555–560.
[32] D. Szelogowski, "SMFSA-Database-And-Form-NN," Apr. 2022.[Online]. Available: https://github.com/danielathome19/Form-NN
[33] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proc. 10th Int. Soc. Music Inf. Retrieval Conf.,* Kobe, Japan, Oct. 2009, pp. 231–236.
[34] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," in *Proc. 8th Int. Soc. Music Inf. Retrieval Conf.*, Vienna, Austria, Sep. 2007, pp. 51–54.
[35] N. Jiang and M. Müller, "Automated methods for analyzing music recordings in sonata form," in *Proc. 13th Int. Soc. Music Inf. Retrieval Conf*., Curitiba, Brazil, Nov. 2013, pp. 595–600.
[36] J. Paulus and A. Klapuri, "Music structure analysis using a probabilistic fitness measure and a greedy search algorithm," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 17, no. 6, pp. 1159–1170, Aug. 2009.
[37] O. Nieto and J. P. Bello, "Music segment similarity using 2D-Fourier magnitude coefficients," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.,* 2014, pp. 664–668.
[38] G. Peeters, "Self-similarity-based and novelty-based loss for music structure analysis," in *Proc. 24th ISMIR Conf*., Milan, Italy, Nov. 5–9, 2023, pp. 749–756.
[39] F. Kaiser, T. Sikora, and G. Peeters, "MIRE X 2012-music structural segmentation task: IRCAMstructure submission," in *Proc. Music Inf. Retrieval Eval. Exchange*, Porto, Portugal, Oct. 8–12, 2012.
[40] C. Cannam et al., "Mirex 2015: VAMP plugins from the centre for digital music," *in Proc. Music Inf. Retrieval Eval. Exchange*, Malaga, Spain, Oct. 26–30, 2015.
[41] J. Salamon, O. Nieto, and N. Bryan, "Deep embeddings and section fusion improve music segmentation," in *Proc. 22nd Int. Soc. Music Inf. Retrieval Conf*., Nov. 2021, pp. 594–601.
[42] J.-C. Wang, J. B. L. Smith, W.-T. Lu, and X. Song, "Supervised metric learning for music structure features," in *Proc. 22nd ISMIR Conf*., Nov. 7–12, 2021, pp. 730–737.
[43] O. Nieto, "MIREX: MSAF v0. 1.0 submission," in *Proc. Music Inf. Retrieval Eval. Exchange*, New York, NY, USA, Aug. 7–11, 2016.
[44] J. de Berardinis, M. Vamvakaris, A. Cangelosi, and E. Coutinho, "Unveiling the hierarchical structure of music by multi-resolution community detection," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 3, no. 1, pp. 82–97,

Jun. 2020.
[45] C. H. Oliván, J. R. B. Blázquez, and D. D.-G. Aparicio, “Music boundary detection using convolutional neural networks: A comparative analysis of combined input features,” *Int. J. Interactive Multimedia Artif. Intell*., vol. 7, pp. 78–88, Dec. 2021.
[46] Y. LeCun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision,” in *Proc. IEEE Int. Symp. Circuits Syst.,* Paris, France, May 30–Jun. 2, 2010, pp. 253–256.
[47] K. Ullrich, J. Schlüter, and T. Grill, “Boundary detection in music structure analysis using convolutional neural networks,” in *Proc.15th Int. Soc. Music Inf. Retrieval Conf.,* Taipei, Taiwan, Oct. 27–31, 2014, pp. 417–422.
[48] T. O’Brien, “Musical structure segmentation with convolutional neural networks,” in *Proc. 17th Int. Soc. Music Inf. Retrieval Conf*., New York, NY, USA, Aug. 7–11, 2016.
[49] T. Grill and J. Schluter, “Music boundary detection using neural networks on spectrograms and self-similarity lag matrices,” in *Proc. 23rd Eur. Signal Process. Conf*., Nice, France, Aug. 31–Sep. 4, 2015, pp. 1296–1300.
[50] T. Grill and J. Schlüter, “Music boundary detection using neural networks on combined features and two-level annotations,” in *Proc. 16th Int. Soc. Music Inf. Retrieval Conf*., Malaga, Spain, Oct. 26–30, 2015, pp. 531–537.
[51] P. Allegraud et al., “Learning sonata form structure on Mozart’s string quartets,” *Trans. Int. Soc. Music Inf. Retrieval Conf*., vol. 2, no. 1, pp. 82–96, Dec. 2019.
[52] M. Gotham and M. Ireland, “Taking form: A representation standard, conversion code, and example corpus for recording, visualizing, and studying analyses of musical form,” in *Proc. 20th Int. Soc. Music Inf. Retrieval Conf.*, Delft, The Netherlands, Nov. 4–8, 2019, pp. 693–699.
[53] C. Weiß, S. Klauk, M. Gotham, M. Müller, and R. Kleinertz, “Discourse not dualism: An interdisciplinary dialogue on sonata form in Beethoven’s early piano sonatas,” in *Proc. 21st Int. Soc. Music Inf. Retrieval Conf.*, Montréal, QC, Canada, Oct. 11–16, 2020, pp. 199–206.
[54] J. Zeitler, C. Weiß, V. ArifiMüller, and M. Müller, “BPSD: A coherent multiversion dataset for analyzing the first movements of Beethoven’s piano sonatas,” *Trans. Int. Soc. Music Inf. Retrieval Conf*., vol. 7, no. 1, pp. 195–212, 2024.
[55] B. H. Van Boer*, Historical Dictionary of Music of the Classical Period*. New York, NY, USA: Scarecrow Press, 2012.
[56] A. Burton, *A Performer’s Guide to Music of the Classical Period*. London, U.K.: Assoc. Board Roy. Sch. Music, 2002.
[57] D. Heartz and B. A. Brown, “Classical,” in *Grove Music Online*, L. Macy, Ed. Oxford, U.K.: Oxford Univ. Press, Jan. 2001, Accessed: Aug. 9, 2025. [Online]. Available: https://doi.org/10.1093/gmo/9781561592630.article.05889
[58] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[59] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, “High-resolution piano transcription with pedals by regressing onset and offset times,” *IEEE/ACM Trans. Audio, Speech, Lang. Process*., vol. 29, pp. 3707–3717, 2021.
[60] C. Hawthorne et al., “Enabling factorized piano music modeling and generation with the MAESTRO dataset,” in *Proc. Int. Conf. Learn. Representations,* New Orleans, LA, USA, May 6–9, 2019.
[61] B. McFee et al., “librosa: Audio and music signal analysis in Python,” in *Proc. 14th Python Sci. Conf.,* Austin, TX, USA, Jul. 6–12, 2015, pp. 18–24.
[62] S. S. Stevens, J. E. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” *J. Acoustical Soc. Amer*., vol. 8, pp. 185–190, Jan. 1937.
[63] C. Harte, M. Sandler, and M. Gasser, “Detecting harmonic change in musical audio,” in *Proc. ACM Int. Multimedia Conf. Exhib*., Santa Barbara, CA, USA, Oct. 23–27, 2006, pp. 21–26.
[64] P. Grosche, M. Müller, and F. Kurth, “Cyclic tempogram—A mid-level tempo representation for music signals,” in *Proc. IEEE ICASSP,* Dallas, TX, USA, Mar. 15–19, 2010, pp. 5522–5525.
[65] A. Clark, “Pillow (PIL fork) documentation,” 2015. [Online]. Available: https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf
[66] J. Foote, “Visualizing music and audio using self-similarity,” in *Proc. 7th ACM Int. Conf*. On Multimedia (Part 1), Orlando, FL, USA, ACM, Oct. 30–Nov. 5, 1999, pp. 77–80.
[67] I. Goodfellow, Y. Bengio, and A. Courville, “Convolutional networks,” in *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, pp. 326–366.
[68] R. Wang, “Edge detection using convolutional neural network,” in *Proc. Int. Symp. Neural Netw.,*

Petersburg, Russia: Springer, Jul. 6–8, 2016, pp. 12–20.
[69] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Representations,* May 7–9, 2015.
[70] Y. Chen, L. Li, W. Li, Q. Guo, Z. Du, and Z. Xu, “Fundamentals of neural networks,” in *AI Computing Systems.* Burlington, MA, USA: Morgan Kaufmann, 2024, pp. 17–51.
[71] F. Chollet et al., “Keras,” 2015. [Online]. Available: https://github.com/fchollet/keras
[72] K. Sechidis, G. Tsoumakas, and I. Vlahavas, “On the stratification of multi-label data,” in *Proc. Mach. Learn. Knowl. Discov. Database*s, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds., Athens, Greece, Sep. 5–9, 2011, pp. 145–158.
[73] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining multi-label data,” in *Data Mining and Knowledge Discovery Handboo*k. Boston, MA, USA : Springer, Jul. 2010, pp. 667–685.
[74] G. Velarde, C. Cancino Chacón, D. Meredith, T. Weyde, and M. Grachten, “Convolution-based classification of audio and symbolic representations of music,” *J. New Music Res*., vol. 47, no. 3, pp. 191–205, May 2018.
[75] F. Wilcoxon, “Individual comparisons by ranking methods,” in *Breakthroughs in Statistics: Methodology and Distribution.* Berlin, Germany: Springer, 1992, pp. 196–202.
[76] C. Nadeau and Y. Bengio, “Inference for the generalization error,” *Mach. Learn*., vol. 52, pp. 239–281, Jan. 2003.
[77] W. S. Gosset, “The probable error of a mean,” *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908.