acoustics

MDPI

Article

SECD: A String Ensemble Chords Dataset for Multi-Task Audio Classification

Angelos Geroulanos *, Panagiotis Zervas and Giannis Tzimas

Department of Electrical and Computer Engineering, University of Peloponnese, 263 34 Patras, Greece; p.zervas@uop.gr (P.Z.); tzimas@uop.gr (G.T.)
* Correspondence: aggelosger@gmail.com or a.geroulanos@go.uop.gr

Abstract

We introduce the String Ensemble Chords Dataset (SECD), a large-scale controlled compositional audio dataset comprising 287,088 harmonic-interval and chord instances constructed through additive superposition of professionally recorded isolated string notes from the Philharmonia Orchestra into duo-, trio-, and quartet-like four-voice mixtures. Each mixture includes exact per-voice metadata for absolute pitch, dynamic marking, and playing technique, and the corpus is organised into six dataset groups covering harmonic intervals, triads, and seventh chords under loose and strict metadata-consistency conditions. To demonstrate dataset utility, we define four representative and reproducible reference benchmarks: ensemble size recognition, triad chord quality identification, per-instrument dynamics classification, and playing technique-family recognition. Baseline Audio Spectrogram Transformer (AST) models achieve test accuracies of 98.67%, 93.73%, 98.19%, and 99.39%, with corresponding macro-F1 scores of 98.64%, 93.73%, 98.01%, and 97.29%, under a complete-instance-disjoint, in-domain evaluation protocol. These results provide reproducible reference performance for the selected SECD tasks and demonstrate the corpus's utility for controlled analysis of harmonic, timbral, dynamic, and textural attributes in classical string audio. The full SECD corpus is released through Zenodo as constructed audio mixtures with accompanying metadata, while the project GitHub repository provides the EXP1–EXP4 benchmark code, saved split definitions, and the mini-SECD demonstration package for lightweight reproducibility.

Keywords: music information retrieval; audio classification; string instruments; chord recognition; dataset; audio spectrogram transformer; transfer learning; playing technique; dynamics; in-domain evaluation

Check for updates

Received: 1 June 2026
Revised: 28 June 2026
Accepted: 2 July 2026
Published: 7 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The automated analysis of Western classical ensemble music presents a set of acoustically challenging problems that remain insufficiently addressed in the Music Information Retrieval (MIR) literature. Unlike isolated instrument recognition or single-voice pitch estimation, ensemble audio combines the timbral signatures of multiple co-occurring instruments into a single waveform, requires sensitivity to per-instrument expressive attributes, and demands that harmonic structure be analysed together with performance-related variation. Progress in this area has been limited not only by the difficulty of the signal-processing problem, but also by the scarcity of large-scale, finely annotated datasets targeting this acoustic domain.

In chamber string performance, polyphonic texture is musically meaningful because the sounding result emerges from interactions among simultaneous instrumental voices rather than from the presence of independent notes alone. Studies of string quartet performance show that ensemble playing involves interdependence across timing, intonation, dynamics, timbre, and articulation, with performers continuously adapting to one another during performance [1–3]. Perception studies likewise show that instrumental combinations may fuse, segregate, or produce emergent timbral identities depending on spectral and temporal relationships among the voices [4]. The SECD is not intended to reproduce this full interactional layer of live chamber performance. It is instead a controlled compositional audio dataset constructed through additive superposition of professionally recorded isolated string notes, representing interval and chord events as reproducible polyphonic audio objects with exact per-voice ground truth.

Existing MIR resources address adjacent problems but do not provide a controlled benchmark for polyphonic string chord analysis with per-voice expressive metadata. Datasets such as the RWC Music Database [5] and MedleyDB [6] provide recordings of complete musical pieces or multitrack material, but their annotations are not designed for per-instrument dynamic and technique attributes within isolated chord events. The IRMAS dataset [7] targets instrument recognition from real recordings, while URMP [8] provides score-aligned multitrack ensemble audio across a heterogeneous instrument set. These resources are valuable, but they do not jointly provide controlled harmonic vocabularies, string-only chord mixtures, per-voice pitch, dynamics, and playing-technique metadata, and reproducible benchmark tasks for ensemble size, chord quality, dynamics, and technique-family recognition.

This gap is important in the current phase of MIR research. Deep learning architectures have substantially advanced audio classification and representation learning, but their long-term scientific value depends on reusable annotated corpora as much as on model design. Peeters argues that annotated datasets are among the most sustainable components of MIR research because they remain usable across successive modelling paradigms [9]. This is especially relevant for fine-grained instrumental analysis. Conventional musical instrument recognition under ordinary playing conditions has been studied extensively, whereas instrumental playing technique recognition remains comparatively underdeveloped and has been identified as a major next step for the field [10]. Existing resources for extended techniques, such as OrchideaSOL, provide detailed isolated-note material, but remain monophonic at the level of the annotated audio event [11]. The SECD addresses this complementary gap by providing polyphonic string interval and chord events annotated not only by harmonic structure, but also by ensemble size, instrument identity, pitch, dynamics, and playing technique.

Recent advances in transformer-based and self-supervised audio representations further motivate a controlled benchmark for this domain. Recent work on learning from audio mixtures has highlighted that the performance of self-supervised representations under polyphonic conditions involving multiple overlapping sound sources remains an important evaluation problem [12]. The Audio Spectrogram Transformer (AST) [13], based on the transformer architecture [14] and related to the Vision Transformer formulation [15], is pre-trained on the AudioSet ontology [16] and has shown strong transfer properties across audio classification tasks. However, such representations have not been systematically evaluated on the specific challenges of polyphonic classical string audio, where harmonic, timbral, dynamic, and voice-multiplicity information coexist in a single mixed signal.

A further motivation for an audio-domain dataset is that notation and chord labels do not fully determine the acoustic event. The score provides an abstract representation of musical structure, but the realised sound depends on performance-mediated factors such

as bowing, articulation, dynamic shaping, timbral control, and ensemble balance [1,17]. Consequently, chord labels such as "major", "minor", or "diminished" are treated here as controlled nominal targets attached to realised audio, rather than as complete descriptions of the perceptual or performance content of the sound.

This paper makes the following contributions:

1. We release the **String Ensemble Chords Dataset (SECD)**, comprising 287,088 annotated harmonic-interval and chord instances constructed through additive superposition of professionally recorded isolated string notes from the Philharmonia Orchestra into controlled duo-, trio-, and quartet-like four-voice configurations (Section 3).
2. We establish **four representative and reproducible reference benchmark tasks** for ensemble size recognition, chord quality identification, dynamics classification, and playing technique recognition, with deterministic train/validation/test splits and reported baseline results (Section 5).
3. We characterise the **controlled source-note reuse** inherent in compositional datasets constructed from finite pools of isolated recordings, quantify its extent in EXP2, and report a complementary EXP2 source-note-disjoint diagnostic with zero exact source overlap between the retained development pool and the diagnostic test set (Sections 4 and 5.3).
4. We provide a **reproducibility package** through the project GitHub repository, including mini-SECD metadata-compatible placeholder WAV files, pre-computed mel spectrogram tensors, benchmark scripts, saved split definitions, and a demonstration notebook for validating the benchmark code path on CPU hardware (Section 8).

2. Related Work

2.1. MIR Datasets and Benchmarks

Annotated audio corpora have been central to progress in MIR. The RWC Music Database [5] established an early standard for structured music collections with genre, beat, and chord labels. MedleyDB [6] advanced the field by providing stem-level multitrack recordings with instrument activity annotations, enabling source separation and predominant instrument recognition research. More recently, FSD50K [18] provided a large-scale general-purpose sound event dataset based on AudioSet categories, while URMP [8] offered score-aligned recordings of small ensembles with per-instrument audio tracks. The SECD differs from these corpora by combining controlled compositional construction for large-scale corpus generation, fine-grained per-instrument expression metadata, and an explicit focus on chord-level analysis in string-ensemble audio.

2.2. Chord Recognition

Automatic chord recognition from audio has been studied extensively, particularly for popular music, where systems typically label continuous audio streams with chord symbols. Deep learning approaches, including Korzeniowski and Widmer's deep chroma extractor [19] and subsequent work incorporating duration models [20], have achieved strong results on popular music chord datasets. Reviews of the broader chord recognition landscape are provided by McVicar et al. [21].

The chord-quality problem addressed in the SECD differs from this setting in three respects. First, samples are isolated chord events rather than continuous audio streams. Second, the instrument set is restricted to bowed strings. Third, the harmonic vocabulary includes diminished and augmented triads as well as seventh-chord qualities, including the half-diminished seventh. Large-vocabulary chord recognition remains difficult because extended chord types may differ by subtle chromatic alterations and because real-world music corpora often exhibit long-tailed chord distributions [22]. Recent work has therefore moved

toward sequence-aware and attention-based architectures, including the Bi-directional Transformer for Chord Recognition and ChordFormer [22,23].

The SECD is complementary to this line of work. Rather than addressing frame-level chord recognition in complete musical recordings, it provides controlled polyphonic audio events with explicit annotations for harmonic interval type, triad quality, seventh-chord quality, instrumentation, voicing, pitch, dynamics, and playing technique. This makes it suitable for evaluating harmonic structure under controlled string-ensemble conditions while separating the task from score alignment, segmentation, key estimation, and full-song harmonic tracking.

2.3. Ensemble Size and Voice-Multiplicity Recognition

Compared with instrument recognition and chord recognition, automatic ensemble size classification has received relatively limited attention in MIR. Grollmisch et al. [24] studied ensemble size classification in Colombian Andean string music recordings, formulating the problem as the estimation of the number of instruments present in an audio mixture. This line of work is relevant to the SECD because it treats voice multiplicity as an audio-inference problem rather than as metadata available from a score or recording description.

The SECD differs from this prior setting in both construction and scope. Whereas Grollmisch et al. focus on real recordings from a specific regional string tradition, the SECD provides controlled harmonic-interval and chord events across duo-, trio-, and quartet-like string configurations. In EXP1, ensemble size recognition is formalised as a three-class benchmark over monaural mixture signals, allowing systematic evaluation of whether a model can infer the number of simultaneously sounding string voices under controlled harmonic, registral, and timbral conditions.

2.4. Harmony, Pitch Combination, and Perceptual Ambiguity

Although chord recognition is often formalised as symbolic label estimation, the perception of harmonic sonorities is shaped by more than abstract interval content. Studies of pitch combinations show that chord perception depends on the arrangement of component tones, register, tonal context, memory, and perceptual grouping [25,26]. Simultaneous intervals are also influenced by sensory consonance, roughness, harmonicity, fusion, musical training, and enculturation [26].

This literature provides context for interpreting the SECD's chord-quality benchmark. The task does not assume that a chord label fully determines perceptual harmony. Instead, it provides a controlled setting in which nominal harmonic categories can be studied as realised string audio under fixed instrumentation, duration, and metadata conditions.

2.5. Instrument Recognition and Playing Technique Classification

Predominant instrument recognition from polyphonic audio [27,28] has attracted sustained attention, with convolutional neural networks achieving strong performance on the IRMAS benchmark [29]. However, identifying the instrument category is a coarser problem than identifying how the instrument is played. Instrumental playing techniques (IPTs) modify the spectral, temporal, and articulatory properties of instrumental sound and are central to musical expressivity. Lostanlen et al. [10] argue that extended playing techniques constitute a major next step for musical instrument recognition because ordinary instrument recognition does not address the fine-grained sound-production categories used in performance practice.

Recent work has continued to emphasise the difficulty of instrumental playing technique detection. Labelled data are scarce, and technique classes are often highly imbalanced because ordinary playing styles occur more frequently than specialised techniques. MERTech addresses these limitations using a self-supervised pre-trained model with multi-

task fine-tuning for instrumental playing technique detection [30]. Other work on bowing-gesture classification in violin performance has shown that bowing-related performance categories can be discriminated using machine-learning methods applied to performer motion data, although such studies typically remain limited to solo-instrument settings and do not address polyphonic audio mixtures [31].

The SECD extends this line of work to polyphonic string mixtures. In EXP4, each present instrument voice is assigned to one of four broad sound-production families: bowed, col legno, harmonic, or plucked. This family-level formulation is more conservative than 21-class atomic technique recognition and reflects the severe imbalance of the underlying technique labels while preserving musically meaningful distinctions between major sound-production regimes in string performance. The complete per-voice metadata retain the original fine-grained technique labels, enabling future work on atomic or hierarchical playing-technique classification without requiring changes to the dataset ontology.

2.6. Timbre, Dynamics, and Ensemble Perception

The classification tasks defined in the SECD are grounded in perceptual dimensions central to musical listening. Timbre is commonly treated as a multidimensional perceptual attribute that cannot be reduced to pitch, loudness, duration, or spatial position, and it plays a central role in source identification, auditory grouping, and instrument tracking [4]. This is particularly relevant for polyphonic string textures, where individual voices may remain perceptually segregated, partially blend, or contribute to an emergent composite sonority.

Dynamics are likewise not reducible to scalar amplitude. Changes in playing level can alter spectral properties such as spectral centroid and spectral spread, meaning that perceived dynamics depend jointly on loudness and timbral change [4]. This observation motivates the EXP3 dynamics task, where dynamic categories are inferred from spectro-temporal patterns rather than from global signal energy alone.

Playing techniques also correspond to acoustically meaningful sound-production categories. In string performance, bowing and articulation choices shape the acoustic signal and are part of the performance-mediated transformation from score to sound [1,17]. The SECD's technique-family benchmark therefore targets broad sound-production regimes while avoiding the stronger claim that technique labels exhaust the expressive meaning of a performance.

This distinction also positions the SECD relative to work on real ensemble performance. Papiotis et al. [3] model string quartet performance as an interdependent process in which timing, intonation, dynamics, timbre, and articulation are coupled across performers. Multimodal quartet-performance resources such as the QUARTET dataset [32] address this ecological interaction layer. The SECD targets a complementary problem: controlled chord-level audio classification with systematic per-voice metadata.

2.7. Transformer-Based Audio Representations

The Audio Spectrogram Transformer [13] adapts the Vision Transformer architecture [15] to audio by treating mel spectrograms as sequences of two-dimensional patches. Pre-trained on AudioSet [16] with 527 audio event categories, the AST has achieved state-of-the-art performance on multiple audio classification benchmarks and demonstrated strong transfer learning properties across diverse acoustic domains. Research on transfer learning and representation learning has shown that models pre-trained on large-scale audio collections can capture spectro-temporal features relevant to music analysis [33–35]. More recently, MERT demonstrated that music-specific self-supervised pre-training can support a broad range of downstream music-understanding tasks [36]. The SECD provides a systematic

evaluation of AST-based transfer learning on the fine-grained string ensemble classification tasks defined in this work.

2.8. Comparison with Existing Datasets

Table 1 positions the SECD against representative existing datasets across the dimensions most relevant to the tasks it supports. While several corpora provide multitrack or multi-instrument recordings, we are not aware of an existing corpus that combines the scale, per-instrument expression annotations, and controlled interval/chord vocabulary required for the four benchmark tasks defined in this work.

The NSynth dataset [37] is the closest precedent in terms of construction methodology: it also uses isolated note recordings to build a large-scale synthesis corpus, but targets single-note timbre modelling rather than polyphonic chord analysis, and does not encode dynamics or playing technique as explicit target labels. OrchideaSOL [11] provides a comprehensive catalogue of orchestral extended techniques, but at the level of individual notes rather than ensemble chords. URMP [8] provides score-aligned ensemble recordings with per-instrument tracks, enabling separation and transcription research, but covers a heterogeneous instrument set with limited chord-level annotation and no explicit dynamics or technique labelling. MedleyDB [6] and the RWC Music Database [5] focus on complete musical pieces at the track or section level, making them unsuitable for the isolated chord classification tasks addressed here.

The SECD is distinctive in simultaneously providing: (1) polyphonic interval and chord events at scale (287K instances); (2) per-instrument dynamic and technique metadata in every sample; (3) a controlled interval/chord vocabulary spanning three complexity levels; and (4) deterministic, reproducible experimental splits for the benchmark tasks. Together, these properties enable a class of classification benchmarks that, to our knowledge, is not jointly supported by an existing corpus.

Table 1. Comparison of SECD with representative related datasets. ✓ = directly supported; ∼ = partially supported; — = not available.

Dataset	Domain	Scale	Synth	Chord	Dyn.	Tech.	Per-Inst.	Strings
SECD (ours)	Strings/chords	287K clips	Yes	✓	✓	✓	✓	✓
NSynth [37]	Multi-inst.	305K notes	Yes	—	∼	∼	—	∼
OrchideaSOL [11]	Orchestral	13K notes	No	—	✓	✓	✓	✓
URMP [8]	Multi-inst.	44 pieces	No	—	—	—	✓	∼
MedleyDB [6]	Pop/Jazz	122 tracks	No	—	—	—	✓	—
RWC [5]	Classical/Pop	315 pieces	No	∼	—	—	—	∼
IRMAS [7]	Multi-genre	6.7K clips	No	—	—	—	—	∼
Bach10 [38]	Baroque str.	10 pieces	No	—	—	—	✓	✓

3. Dataset Construction

3.1. Source Material and Acoustic Basis

The SECD is constructed from the Philharmonia Orchestra Sound Samples library [39], a professionally recorded collection of isolated notes covering a broad range of pitches, dynamics, and playing techniques for orchestral instruments. The released SECD corpus does not redistribute the original isolated solo recordings. Instead, it contains controlled compositional audio mixtures created through additive superposition of selected string-note recordings into harmonic-interval and chord instances.

The mixture construction uses four string voice roles: cello, viola, violin 1, and violin 2. The violin 2 voice is not an independently recorded source instrument; it is implemented as a spectrally differentiated derivative of the violin source recordings via a fixed, mild transformation consisting of -0.15 semitone detuning, low-frequency emphasis, and high-frequency attenuation. This procedure differentiates the two violin roles and avoids exact

waveform duplication in quartet-like textures while preserving the nominal pitch label and musical voice function. The resulting violin 2 voice is a fixed role-specific component of the SECD construction procedure, rather than a full acoustic or performative model of independently recorded second-violin performance. The present benchmark design does not isolate the individual contribution of this transformation from instrument role, sample support, quartet-only context, or the overall mixture configuration.

A corpus-level property inherited from the source material is the predominance of standard bowed string production, especially `arco-normal`. From an acoustical perspective, sustained bowed-string tone production is closely associated with stable Helmholtz motion, the stick–slip vibration regime underlying the characteristic full tone of bowed string instruments [40,41]. From a performance-practice perspective, `arco-normal` represents the default sound-production mode of orchestral bowed strings, whereas techniques such as col legno, pizzicato, harmonics, sul ponticello, and related articulations function as more specialised or colouristic resources [41,42].

This musical and acoustical asymmetry is also reflected at the corpus-construction level. The initial pool of professionally recorded solo notes contains substantially more standard bowed material than special-technique material. Because the SECD is generated by combinatorially superposing source-note recordings, this source-level distribution is inherited by the derived interval/chord corpus and amplified at the instance and pooled voice-instance levels. Consequently, the dominance of `arco-normal` and, after family-level mapping, of the broader bowed class is an expected corpus property rather than an uncontrolled annotation bias.

3.1.1. Rationale for String-Only Instrumentation

The Philharmonia Orchestra Sound Samples library provides a broader instrumental palette than the one used in the SECD, including standard orchestral instruments as well as guitar, mandolin, banjo, and percussion [39]. The restriction to cello, viola, violin 1, and violin 2 is therefore a design decision rather than a limitation of the source material.

The first motivation is musical and structural. Violin, viola, and cello belong to a historically established chamber-music domain in which duo, trio, and quartet configurations provide canonical models of polyphonic string texture. Focusing on this ensemble tradition allows SECD to model chordal string writing in a way that remains close to mainstream chamber repertoire while avoiding the broader and stylistically heterogeneous setting of full-orchestra writing.

The second motivation is perceptual. Because these instruments belong to the same bowed-string family and occupy partially overlapping registral and timbral regions, their mixtures create conditions in which source identity, voice separation, and the perceived number of simultaneous voices are non-trivial. This formulation is consistent with timbre-perception research showing that timbre supports source identification and auditory grouping, while instrumental combinations may segregate, blend, fuse, or form emergent composite sonorities [4]. It is also consistent with interval and polyphony research showing that simultaneous tones may fuse perceptually and that musical-interval perception is affected by register, timbre, context, and grouping mechanisms [26].

The third motivation is methodological. Holding the instrument family constant allows SECD to vary ensemble size, chord type, dynamics, and playing technique while reducing the risk that broad cross-family timbral contrasts dominate the classification tasks. SECD therefore follows a controlled experimental strategy: complexity is introduced along selected musical dimensions while the global instrumental family is held fixed. The spectrally differentiated violin 2 part should be understood as a pragmatic device

for constructing quartet-like four-voice string textures, not as a full acoustic model of second-violin performance practice [1,2].

Each individual note recording is preprocessed to a uniform format: mono, 22.05 kHz sampling rate, and 2.0 s duration by trimming or zero-padding. Phrase recordings are excluded to retain isolated note events. Harmonic interval and chord waveforms are then constructed by linear superposition of the constituent solo note recordings:

$$x(t) = \sum_{i=1}^{N} s_i(t), \tag{1}$$

where $N \in \{2,3,4\}$ denotes the ensemble size and $s_i(t)$ is the i-th instrument's note waveform. After superposition, peak-based rescaling is applied where necessary to prevent 16-bit integer overflow.

This rescaling is relevant to the interpretation of the dynamics experiment. Because mixture-level peak rescaling can partially reduce absolute amplitude differences between samples, EXP3 should not be interpreted as waveform-amplitude recovery alone. Dynamic labels may also be encoded through relative energy distribution, spectral envelope, instrument-specific response, and timbral changes associated with playing level.

The additive construction provides exact ground truth for all per-instrument attributes, but it does not include the mutual acoustic coupling and performer-interaction effects present in live ensemble recordings. In live string ensemble performance, musicians adapt to one another through auditory, visual, and gestural feedback, shaping timing, intonation, dynamics, timbre, and articulation as part of an interdependent process [3]. SECD therefore targets controlled construction of polyphonic string-chord audio with exact per-voice metadata, rather than ecological simulation of live quartet performance.

3.1.2. Instrument–Pitch Ordering Constraint

All interval and chord instances are constructed under a fixed instrument–pitch ordering constraint. For every generated sample, pitches are assigned in strictly ascending pitch order across instruments, enforcing a deterministic mapping between pitch height and instrument role.

For two-instrument configurations, the lower pitch is assigned to the lower-register instrument in the selected pair and the higher pitch to the upper-register instrument. For three-instrument configurations, the lowest pitch is assigned to cello, the next to viola, and the highest to violin 1. For four-instrument configurations, the ordering is extended such that cello carries the lowest pitch, followed by viola, violin 2, and violin 1 as the highest voice.

This constraint eliminates permutation-equivalent realisations of the same harmonic structure and ensures consistent voicing across the dataset. Consequently, harmonic intervals, triads, and seventh chords are represented in a canonical register-distributed form rather than as arbitrary permutations of the same pitch classes. The design is informed by conventional register-based role distribution in tonal string writing and orchestration, where lower instruments typically realise bass functions, upper instruments carry higher voices, and inner voices mediate the harmonic space between them [42–44]. This canonical setting does not cover voice crossing, alternative distributions of chord tones and registers among the instruments, or the full range of registral layouts encountered in polyphonic string writing.

The constraint is especially important for chord inversions. In the SECD, inversion is not treated as a purely symbolic label detached from the audio signal. Because the lowest pitch is assigned to the cello in triadic and seventh-chord textures, inversion determines which chord member occupies the bass register and therefore changes the registral layout,

spectral weighting, and perceived harmonic profile of the chord. This makes inversion metadata acoustically meaningful and suitable for future inversion-aware recognition experiments, even though no dedicated inversion-recognition benchmark is reported in the present paper.

From a perceptual perspective, the fixed register-distributed design also reduces uncontrolled ambiguity in polyphonic texture formation. Register separation and timbral role assignment support auditory organisation of simultaneous instrumental lines while still allowing chord tones to interact through masking, fusion, and spectral overlap [4,25,26]. The SECD therefore follows the logic of controlled stimulus construction: it omits some properties of live ensemble performance, including room interaction, within-ensemble adaptation, and expressive timing, in exchange for exact per-instrument ground truth and systematic coverage of harmonic interval, chord, dynamic, and technique combinations.

3.2. Filename Encoding and Metadata Schema

Each source note file encodes four attributes in its filename: voice role, pitch, dynamic marking, and playing technique. Formally, the source-level metadata associated with source note i is the tuple

$$m_i = (\text{role}_i,\ p_i,\ d_i,\ t_i), \tag{2}$$

where $\text{role}_i \in \{\text{cello}, \text{viola}, \text{violin1}, \text{violin2}\}$, p_i is the encoded absolute pitch label, d_i is the dynamic marking, and t_i is an atomic playing-technique label encoded directly in the source filename metadata.

These source-level attributes are propagated into the metadata of each constructed SECD mixture. Thus, while the released corpus contains mixture WAV files rather than isolated source-note WAVs, the metadata preserves the voice role, pitch, dynamic, and technique attributes of every present voice in each mixture.

The metadata schema separates nominal musical categories from their acoustic realisation. Dynamic labels, technique labels, interval labels, and chord labels are treated as supervised targets because they are musically meaningful and acoustically consequential, but they should not be read as complete descriptions of performance expression. In particular, dynamics may affect both loudness and spectral shape, while playing-technique labels capture discrete performance instructions rather than the full continuum of expressive bowing behaviour [4,17].

3.3. Dataset Groups and Subset Taxonomy

The SECD is organised into six dataset groups, structured by ensemble size and harmonic complexity. Each group exists in two subset types: a *loose* subset, in which different dynamics and playing techniques may be assigned to different instruments within the same interval or chord instance, and a *strict* subset, in which all instruments in an interval or chord instance share identical dynamics and playing techniques. This metadata-consistency constraint substantially reduces the number of valid configurations and hence corpus size.

- **harmonic_intervals** (duo): Two-instrument combinations spanning 20 interval classes from unison to major thirteenth. The duo subset is constructed across the three instrument pairings cello–viola, cello–violin, and viola–violin, ensuring coverage of all pairwise combinations within the string trio configuration. Loose: 90,000 samples (30,000 per instrument pair); strict: 11,224.
- **triads** (trio): Three-instrument chords with four chord quality labels—major, minor, diminished, and augmented. Loose: 80,000 (20,000 per quality); strict: 4586.

- **7th_chords** (quartet): Four-instrument chords with four types: major seventh (maj7), dominant seventh (7), minor seventh (min7), and half-diminished seventh (m7♭5). Loose: 92,598; strict: 8680.

Aggregate statistics are summarised in Table 2. The full SECD corpus is the union of the loose and strict subsets, yielding 287,088 harmonic-interval and chord instances.

Table 2. SECD statistics by group.

Group	Ensemble	Labels	Loose	Strict
harmonic_intervals	duo	20 intervals	90,000	11,224
triads	trio	4 qualities	80,000	4586
7th_chords	quartet	4 types	92,598	8680
Total			**262,598**	**24,490**

3.4. Full SECD Label Vocabulary and Structural Definition

For the full SECD corpus, the label space is defined by the original file-level metadata and by deterministic mappings derived from it. The SECD comprises 287,088 harmonic interval and chord instances organised into three structurally distinct groups: 101,224 `harmonic_intervals` samples for two-instrument ensembles, 84,586 `triads` samples for three-instrument ensembles, and 101,278 `7th_chords` samples for four-instrument ensembles. The ensemble ontology is fixed by construction: harmonic intervals correspond to duos, triads to trios, and seventh chords to quartet-like four-voice textures.

The resulting full SECD label space is summarised in Table 3.

Table 3. Compact summary of the full SECD label space.

Component	Cardinality	Definition
Harmonic intervals	20	Duo interval classes
Dynamic labels	8	Discrete levels and continuous processes
Playing techniques	21	Filename-encoded atomic technique labels
Triad chord types	4	Major, minor, diminished, augmented
7th chord types	4	7, min7, maj7, m7♭5

3.4.1. Harmonic Interval Vocabulary

The `harmonic_intervals` portion spans 20 interval labels: Unison; Minor 2nd; Major 2nd; Minor 3rd; Major 3rd; Perfect 4th; Tritone; Perfect 5th; Minor 6th; Major 6th; Minor 7th; Major 7th; Octave; Minor 9th (b9); Major 9th (9); Augmented 9th (#9); Perfect 11th (11); Augmented 11th (#11); Minor 13th (b13); and Major 13th (13). The interval ontology includes both simple and compound intervals up to the major 13th. In the full corpus, these 20 interval classes are approximately uniformly distributed.

3.4.2. Dynamic Labels

Dynamic annotation comprises eight labels: pianissimo, piano, mezzo-piano, mezzo-forte, forte, fortissimo, crescendo, and decrescendo. The first six correspond to discrete dynamic levels, whereas crescendo and decrescendo represent continuous dynamic processes. In the full dataset, the continuous dynamic labels are present but substantially more sparsely represented than the discrete dynamic levels.

3.4.3. Playing Techniques

The 21 ground-truth playing technique labels are: arco-normal, arco-sul-ponticello, arco-sul-tasto, arco-au-talon, arco-tremolo, arco-martele, arco-minor-trill, arco-major-trill, arco-glissando, arco-col-legno-battuto, arco-col-legno-tratto, arco-harmonic, pizz-normal,

pizz-glissando, pizz-tremolo, snap-pizz, natural-harmonic, artificial-harmonic, molto-vibrato, non-vibrato, and con-sord. These labels correspond directly to the filename-encoded metadata propagated into the full SECD metadata tables. The technique distribution is highly imbalanced, with `arco-normal` dominating the atomic label space. This imbalance reflects both the musical role of normal bowing as the default string sound-production mode and the structure of the original source-note pool, which contains more standard bowed recordings than special-technique recordings. When these atomic labels are mapped into broader sound-production families, this source-level predominance produces a strong majority for the bowed family.

3.4.4. Chord Types and Inversions

For the trio ensemble, triad quality is annotated with four labels: major, minor, diminished, and augmented. For the quartet ensemble, seventh-chord quality is annotated with four labels: dominant seventh (7), minor seventh (min7), major seventh (maj7), and half-diminished seventh (m7♭5). Inversion is explicitly encoded for both chord classes. Triads include root position, first inversion, and second inversion. Seventh chords include root position, first inversion, second inversion, and third inversion.

This inversion metadata is musically consequential rather than merely descriptive. In tonal harmony, inversion changes the chord member assigned to the bass voice and therefore alters the registral layout, functional interpretation, and acoustic weighting of the chord. In the SECD, this effect is explicit because the cello is always assigned the lowest pitch in triadic and seventh-chord textures; consequently, inversion determines whether the cello carries the root, third, fifth, or, for seventh chords, the seventh of the chord. The dataset therefore supports future inversion-aware recognition tasks in addition to the chord-quality benchmark reported in EXP2.

3.5. Acoustic Feature Representation

For the AST-based benchmark experiments, SECD WAV mixtures are converted into mel spectrogram representations. All interval and chord waveforms are resampled to 16 kHz for feature extraction. Mel spectrograms are computed using 128 mel-frequency bins over a 256-frame temporal window, yielding a fixed-size feature map of shape 128×256, consistent with the input representation used for the AST-based experiments in this work [13].

For compatibility with the pre-trained AST backbone, mel spectrograms are processed using the `ASTFeatureExtractor` preprocessing convention. The extractor computes 128-bin mel spectrogram features from 16 kHz audio, pads or truncates each input to 256 frames, and applies the stored mean and standard-deviation normalisation parameters associated with the AudioSet-pre-trained AST configuration. The resulting fixed-size tensors have shape 128×256 and are cached as NumPy arrays (`.npy`) for computational efficiency during EXP1–EXP4. These cached tensors are derived experimental features and should not be confused with the primary SECD release, which consists of constructed WAV mixture files and metadata.

4. Evaluation Protocol and Data Splitting

4.1. Deterministic Stratified Experimental Splitting

The SECD is distributed as a unified corpus without predefined training, validation, or test partitions. For the benchmark experiments (EXP1–EXP4), each experiment selects a task-specific subset of the SECD and applies deterministic stratified partitioning into 70% training, 15% validation, and 15% test data.

Splitting is performed at the level of complete harmonic interval or chord instances after the task-specific filtering and label mapping required by each experiment. Split indices are saved to disk and reused in subsequent runs to ensure exact reproducibility. This design separates the released dataset organisation from the experimental evaluation protocol: the SECD is released as a unified corpus, whereas train/validation/test partitions are defined within the benchmark pipelines.

The splitting procedure used for the EXP1–EXP4 benchmark experiments is therefore complete-instance-disjoint: no complete mixture instance crosses the training, validation, and test partitions. However, it is not source-note-disjoint, because the underlying solo note recordings may be reused in different mixtures assigned to different partitions.

4.2. Compositional Structure and Source-Note Reuse

A fundamental property of the SECD follows from its construction methodology. All interval and chord instances are generated by combinatorially superposing a finite pool of professionally recorded solo note samples. Consequently, individual source-note recordings may be reused across multiple mixtures and may occur in different partitions under distinct harmonic and structural contexts.

For example, a specific cello note recording (e.g., C3 played with a given dynamic and technique) used within a C-major triad in the training set may also appear as part of a different chord or interval in the validation or test sets. This reuse occurs at the level of atomic source-note recordings, not at the level of complete interval or chord mixtures.

This reuse is a structural consequence of constructing a large-scale controlled compositional dataset from a finite pool of isolated recordings. It enables systematic recombination of musical attributes while preserving exact per-voice metadata across a large number of polyphonic configurations.

To quantify the extent of source-note reuse in the original EXP2 protocol, source identity was defined as the exact stored value of each explicit constituent-source field (`cello_file`, `viola_file`, and `violin1_file`) after removing surrounding whitespace only. No instrument mapping, filename remapping, or inferred source identity was used. Under this definition, the original training partition contained 1422 unique exact source IDs, whereas the test partition contained 1394, of which 1392 were shared. Consequently, 99.86% of the exact source IDs present in the original test partition had already appeared in the training partition. At the mixture level, 12,686 of the 12,688 original test mixtures contained three constituent sources previously observed during training, while the remaining two mixtures contained two previously observed sources. No original test mixture was composed entirely of unseen exact source-note recordings. All exact source IDs occurring in the original validation partition were also present in the training partition.

Using only existing EXP2 mixtures, we additionally constructed a complementary source-note-disjoint diagnostic. A held-out exact-source set was defined, and a mixture was retained in the diagnostic test set only if all three of its constituent source IDs belonged to that set. A mixture was retained in the development pool only if all three source IDs belonged to the non-held-out set. Mixtures combining held-out and non-held-out source IDs were excluded. Because all EXP2 mixtures were connected through shared source-note recordings, a source-note-disjoint partition retaining every mixture was not possible. The retained development pool was then divided into training and validation partitions using a stratified 85/15 split based on the joint label `chord_type` $\times$ `subset`, with random seed 1337. The resulting diagnostic dataset comprised 32,456 training mixtures, 5728 validation mixtures, and 1148 diagnostic test mixtures, while 45,254 cross-partition mixtures were excluded. The diagnostic test set retained all four chord-quality classes and included samples from both the loose and strict subsets. There was zero exact source overlap between

the retained development pool and the diagnostic test set. This substantial reduction in retained mixtures reflects the dense reuse structure of EXP2 and illustrates the practical cost of enforcing exact source-note disjointness in a combinatorially constructed corpus.

This diagnostic protocol evaluates generalisation to previously unseen exact source-note recordings within the controlled SECD domain and complements, rather than replaces, the complete-instance-disjoint EXP2 benchmark.

The implications of source-note reuse differ across tasks. For ensemble-size and chord-quality recognition, no individual source-note recording uniquely determines the target label: the same source note may appear in different ensemble sizes, chord qualities, voicings, and harmonic contexts. For the voice-level dynamics and technique-family tasks, however, the target labels are inherited directly from the source-note metadata. The EXP1–EXP4 benchmark evaluation therefore does not isolate the possible contribution of repeated source-note recordings from that of mixture-level harmonic, spectral, and textural structure. The complementary EXP2 source-note-disjoint diagnostic removes exact source overlap at the development–test boundary and provides a targeted assessment of the sensitivity of chord-quality performance to source reuse. As reported in Section 5.3, the observed reduction in performance indicates that this evaluation regime is substantially more challenging than the complete-instance-disjoint EXP2 benchmark. At the same time, the retained diagnostic performance indicates that triad quality remains partly recoverable from mixtures composed entirely of previously unseen exact source-note recordings within the controlled SECD domain. These findings do not establish either the presence or the absence of shortcut learning, but provide a more precise characterisation of the evaluation boundary for EXP2. The diagnostic should not be interpreted as source-note-disjoint validation of EXP1, EXP3, or EXP4.

Controlled source-note reuse should be distinguished from duplicate-instance leakage, in which identical or overlapping complete recordings cross training and evaluation partitions. Prior work has shown that such overlap can yield overly optimistic performance estimates and that recordings sharing a common recording process may also create weaker forms of contamination, motivating grouped splitting strategies [45]. In both evaluation protocols implemented in this study, no complete SECD mixture crosses the relevant training, validation, and test partitions. Under the EXP1–EXP4 benchmark protocol, however, identical atomic source-note recordings may reappear within different mixtures and across partitions. The complementary EXP2 diagnostic additionally removes this exact source overlap between the retained development pool and the diagnostic test set.

Accordingly, we distinguish three evaluation regimes. Complete-instance-disjoint evaluation holds out complete interval or chord mixtures, although their constituent source-note recordings may have appeared during model development. Source-note-disjoint evaluation additionally requires the exact source-note recordings used in the test set to be absent from the development data. External-recording evaluation instead uses independently recorded ensemble audio outside the controlled SECD construction process. The EXP1–EXP4 benchmark results follow the first regime, whereas the complementary EXP2 diagnostic provides targeted evidence under the second. The third regime remains outside the scope of the present study. Neither of the two implemented protocols should therefore be interpreted as evidence of generalisation to independently recorded ensemble audio.

5. Benchmark Experiments

5.1. Model Architecture and Training Configuration

Across the four SECD benchmark tasks, we use the Audio Spectrogram Transformer (AST) as the shared backbone architecture. AST applies a Transformer encoder directly to sequences of spectrogram patches, following the self-attention architecture introduced

by Vaswani et al. [14] and the patch-based formulation of the Vision Transformer [15]. Gong et al. [13] describe AST as a convolution-free, attention-based model for audio classification and report strong results across several audio classification benchmarks.

This architecture is appropriate for the SECD because the target labels are not determined by isolated local spectral cues alone. Ensemble size, chord quality, dynamics, and technique family may depend on relationships distributed across frequency bands and time frames, especially when multiple string voices overlap within a single monaural mixture. AST therefore provides a high-capacity reference backbone for evaluating whether an attention-based audio representation can support harmonic, timbral, dynamic, and voice-multiplicity classification in controlled polyphonic string audio.

The shared architecture family and its task-specific heads are illustrated in Figure 1.

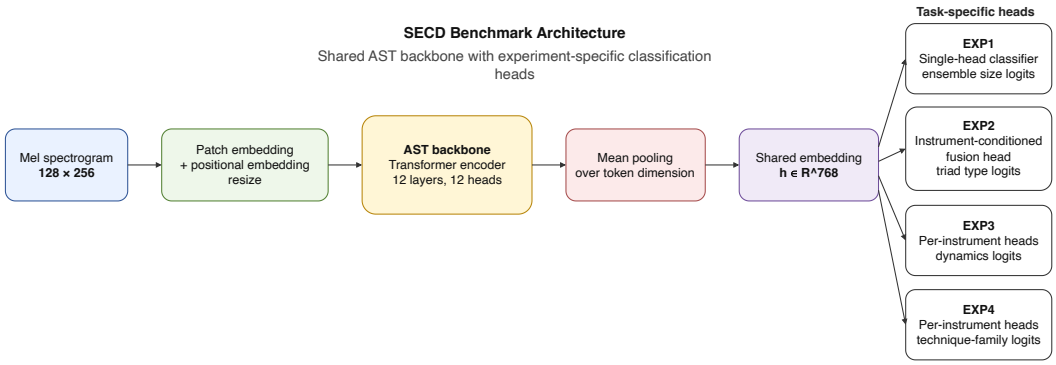

Figure 1. Unified architecture family used for the SECD benchmark tasks. A mel spectrogram input (128×256) is processed by a pre-trained Audio Spectrogram Transformer (AST) backbone, producing a shared embedding $\mathbf{h} \in \mathbb{R}^{768}$. Task-specific heads operate on this embedding: a single-head classifier (EXP1), instrument-conditioned fusion heads (EXP2), and per-instrument multi-head outputs for dynamics and technique-family prediction (EXP3–EXP4).

All four experiments use the same general architecture family: a pre-trained AST backbone followed by task-specific classification heads. Let $\mathbf{X} \in \mathbb{R}^{128\times256}$ denote a normalised mel spectrogram. The AST backbone f_θ maps $\mathbf{X}$ to a sequence of hidden states, which are mean-pooled to obtain a fixed-dimensional embedding:

$$\mathbf{h} = \text{MeanPool}(f_\theta(\mathbf{X})) \in \mathbb{R}^{768}. \tag{3}$$

The backbone consists of 12 transformer encoder blocks with 12 attention heads, 768-dimensional hidden representations, and 3072-dimensional feedforward layers. Since the input spectrograms contain 256 time frames, positional embeddings are resized to match the corresponding patch grid.

The classification stage varies by task. For single-head settings, the classification head g_ϕ maps $\mathbf{h}$ to class logits via a two-layer Multi-Layer Perceptron (MLP):

$$\hat{y} = g_\phi(\mathbf{h}) = W_2\,\text{Dropout}(\text{GELU}(W_1\,\mathbf{h} + b_1)) + b_2, \tag{4}$$

where $W_1 \in \mathbb{R}^{768\times768}$, $W_2 \in \mathbb{R}^{C\times768}$, and Dropout is applied with experiment-dependent probability. C denotes the number of target classes.

For EXP2, instrument-conditioned representations are constructed by adding instrument embeddings $\mathbf{e}_i$ to the shared embedding $\mathbf{h}$ and passing them through instrument-specific projection heads:

$$\mathbf{z}_i = q_i(\mathbf{h} + \mathbf{e}_i), \tag{5}$$

followed by concatenation and fusion into global chord-type logits:

$$\hat{y} = g_\phi([\mathbf{z}_1; \mathbf{z}_2; \mathbf{z}_3]). \tag{6}$$

For EXP3 and EXP4, multi-head classification is used. Each instrument i is associated with a separate prediction head operating on $\mathbf{h} + \mathbf{e}_i$. In EXP3, dynamics labels are predicted:

$$\hat{y}_i^{dyn} = g_i^{dyn}(\mathbf{h} + \mathbf{e}_i), \tag{7}$$

while in EXP4, technique-family labels are predicted:

$$\hat{y}_i^{fam} = g_i^{fam}(\mathbf{h} + \mathbf{e}_i). \tag{8}$$

Absent instruments are masked during loss computation using an ignore index. These reference baselines use a shared mixture representation with instrument-specific prediction heads, but they do not explicitly model structured dependencies among the simultaneous instrumental voices.

The four experiments use a common fine-tuning protocol, with task-specific differences only where required by the target definition and output structure. All experiments fine-tune the pre-trained `MIT/ast-finetuned-audioset-10-10-0.4593` AST checkpoint using fixed-size 128×256 mel-spectrogram tensors. The AST backbone is common to all tasks, while the classification heads differ according to whether the task is global single-label classification (EXP1–EXP2) or pooled per-instrument voice-level classification (EXP3–EXP4).

The training configuration used in the reported runs is summarised in Table 4. Optimisation and scheduling are handled through the Hugging Face `Trainer` API. Evaluation, checkpointing, and logging are performed at epoch level, and the best checkpoint is selected according to validation macro-F1. Parameters not explicitly overridden in the experiment notebooks follow the corresponding `TrainingArguments` defaults.

Table 4. Core training configuration for the four SECD benchmark experiments. Settings are reported from the executed experiment notebooks.

Setting	EXP1	EXP2	EXP3	EXP4
Task	Ensemble size	Triad quality	Dynamics	Technique family
Input tensor	128×256	128×256	128×256	128×256
Backbone checkpoint	`MIT/ast-finetuned-audioset-10-10-0.4593`			
Epochs	40	40	40	40
Learning rate	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}
Warmup ratio	0.05	0.05	0.05	0.05
Weight decay	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Seed	1337	1337	1337	1337
Label smoothing	0.05	0.05	0.05	0.05
Class weighting	effective-number	effective-number	effective-number	effective-number
Best-model metric	macro-F1	macro-F1	macro-F1	macro-F1

Class imbalance is handled through effective-number class weights computed from the corresponding training split. EXP1 and EXP2 use weighted cross-entropy for global single-label classification. EXP3 and EXP4 use masked weighted cross-entropy over per-instrument heads: absent instrument positions are encoded with an ignore index of -1 and excluded from both loss computation and pooled voice-level metric calculation. No weighted sampler is used in the executed experiment notebooks. Table 4 documents the reported benchmark runs rather than defining a mandatory training recipe for future SECD experiments; alternative architectures and optimisation settings can be evaluated using the released splits and metadata.

Representative normalised mel spectrogram inputs spanning the four benchmark settings are shown in Figure 2.

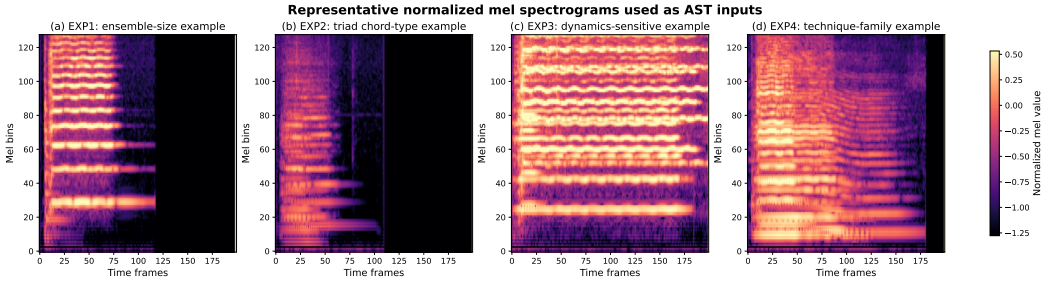

Figure 2. Representative normalised mel spectrograms from SECD subsets, illustrating the time–frequency structure of inputs to the AST-based models across the four experimental settings: (**a**) ensemble size (EXP1), (**b**) triad chord type (EXP2), (**c**) dynamics variation (EXP3), and (**d**) technique-family variation (EXP4).

5.2. EXP1: Ensemble Size Recognition

The first task is a 3-class classification problem: given a harmonic-interval or chord audio clip, the model predicts whether it was produced by a duo-, trio-, or quartet-like four-voice texture. Training data are drawn from the three loose SECD groups: `harmonic_intervals_loose`, `triads_loose`, and `7th_chords_loose`, which map deterministically to the ensemble-size labels {duo, trio, quartet}. The strict subsets are not used in this experiment. The EXP1 corpus contains 262,598 samples: 90,000 duos, 80,000 trios, and 92,598 quartets. A deterministic stratified 70/15/15 split is applied at the complete interval/chord-instance level, yielding 183,818 training samples, 39,390 validation samples, and 39,390 test samples. The held-out test set contains 13,500 duo samples, 12,000 trio samples, and 13,890 quartet samples.

Results on the held-out test set are reported in Table 5. The model achieves an overall accuracy of 98.67% and a macro-averaged F1 score of 98.64%, with consistently high performance across all three ensemble-size classes. The duo class obtains the highest per-class F1 score (99.29%), followed by quartet (98.73%), while trio is marginally the most challenging class (97.90%). To our knowledge, no prior work has reported results on controlled ensemble-size classification for polyphonic string interval/chord audio under this formulation.

Table 5. EXP1 test results: ensemble size recognition.

Class	Support	Precision (%)	Recall (%)	F1 (%)
duo	13,500	99.22	99.36	99.29
trio	12,000	98.12	97.68	97.90
quartet	13,890	98.61	98.86	98.73
Macro avg	39,390	98.65	98.63	**98.64**
Accuracy				**98.67**

Despite the high classification accuracy, the task remains perceptually non-trivial. Increasing the number of simultaneous string voices does not necessarily produce a simple additive increase in independently identifiable sources. Polyphonic listening involves both segregation and fusion: additional voices may be perceived as separate streams, may blend with existing voices, or may contribute to a composite timbre distinct from its individual components [4,26].

The confusion matrix in Figure 3 shows that errors are concentrated between adjacent ensemble sizes, particularly between trio and quartet. This pattern is consistent with the underlying acoustics: the transition from three to four simultaneous voices introduces relatively subtle changes in spectral density and harmonic overlap compared with the more distinct transition from duo to trio. The near-perfect separation of the duo class indicates

that two-voice textures exhibit sufficiently distinct spectro-temporal characteristics under SECD's controlled conditions.

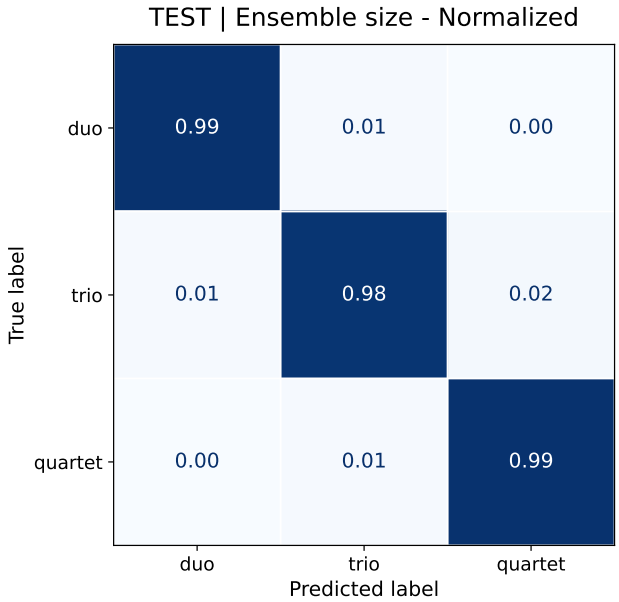

Figure 3. Normalised confusion matrix for EXP1 ensemble size recognition. Minor errors are concentrated between the trio and quartet classes, indicating the acoustic proximity of three- and four-voice textures under controlled voicing and instrumentation conditions. Darker cells indicate higher row-normalised values, while lighter cells indicate lower values.

Although the absolute number of errors remains low, with 181 misclassifications from trio to quartet and 151 from quartet to trio, normalised results are reported to enable direct comparison across classes with different support.

Overall, these results indicate that ensemble size is highly recoverable within the SECD distribution from mel spectrogram representations using a transformer-based model. The findings should be interpreted as complete-instance-disjoint, in-domain reference performance rather than as evidence of generalisation to source-note-disjoint or independently recorded ensemble audio.

5.3. EXP2: Triad Chord Quality Recognition

The second task is a 4-class classification problem: given a chord audio clip, the model predicts its triad quality among {major, minor, diminished, augmented}. Samples are drawn from the full triad portion of the SECD, comprising both the `triads_loose` and `triads_strict` subsets. Chord instances are constructed through the superposition of three independently recorded string voices (cello, viola, violin 1), following the voicing and instrumentation constraints described in Section 3. The EXP2 corpus contains 84,586 triad samples: 80,000 from `triads_loose` and 4586 from `triads_strict`. The corpus is organised across four triad chord-quality classes: major, minor, diminished, and augmented. A deterministic 70/15/15 split is applied at the complete-chord level, stratified by the joint label `chord_type` × `subset`, so that both chord quality and loose/strict membership are preserved across partitions. The resulting split contains 59,210 training samples, 12,688 validation samples, and 12,688 test samples. The held-out test set contains class supports of 3161 major, 3161 minor, 3159 diminished, and 3207 augmented samples.

A central challenge in this task is that major and minor triads differ by a single semitone in one voice. Consequently, the discriminative signal resides in subtle spectro-temporal differences rather than in large-scale structural changes.

Results are reported in Table 6. The model achieves 93.73% accuracy and 93.73% macro-F1. The dominant error mode is major–minor confusion, with normalised off-diagonal entries of approximately 5–6% in the held-out test set (Figure 4). This behaviour is consistent with prior work in chord recognition and auditory perception, where small pitch alterations can yield acoustically similar sonorities while remaining categorically distinct [21].

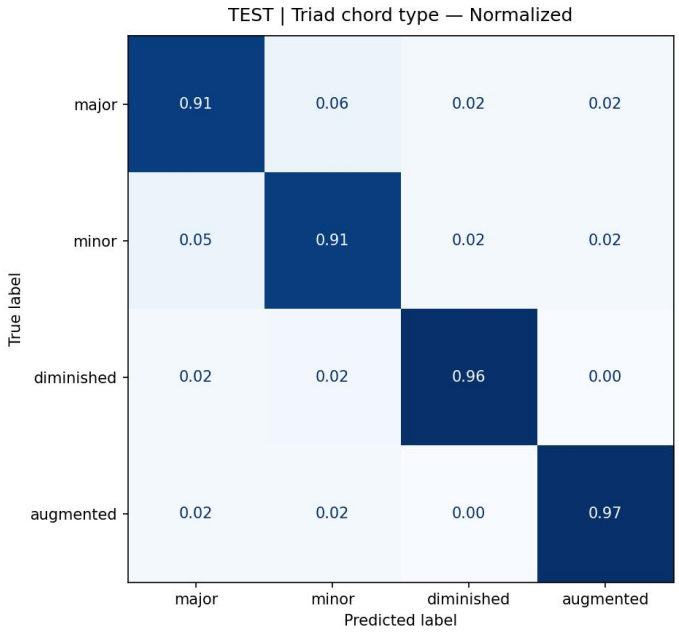

Figure 4. Normalised confusion matrix for EXP2 triad chord quality recognition. The dominant off-diagonal structure appears between major and minor triads (≈5–6%), reflecting their minimal pitch difference and resulting spectral similarity under controlled recorded-source conditions. Diminished and augmented triads exhibit higher separability, with near-zero confusion between them. Darker cells indicate higher row-normalised values, while lighter cells indicate lower values.

From a perceptual standpoint, this pattern is expected. Interval perception is not governed by pitch distance alone, but is modulated by timbre, register, and spectral interaction between simultaneous tones [25,26]. In the SECD, all chords are constructed from professionally recorded isolated instrument notes under controlled conditions. Because EXP2 uses both loose and strict triad subsets, the model is exposed to chord instances with both within-chord metadata variability and fully matched dynamic/technique configurations. The task therefore requires the model to distinguish harmonic structure under timbre- and metadata-dependent variability rather than under a single uniform metadata regime.

In contrast, diminished and augmented triads achieve higher F1 scores, at 96.25% and 96.54%, respectively. A cautious interpretation is that these chord qualities induce more distinctive spectro-temporal patterns under the constrained voicing and instrumentation regime of the SECD. However, this observation should not be generalised to broader musical perception, where chord identification is influenced by tonal context, voice ordering, fusion effects, and listener experience in addition to intervallic structure [25,26].

To complement the complete-instance-disjoint EXP2 benchmark, we evaluated the same AST architecture and training configuration under the source-note-disjoint diagnostic protocol described in Section 4. The diagnostic used the same preprocessing, cached input features, and label definition as the reported EXP2 benchmark, and no diagnostic-specific hyperparameter optimisation was performed. It used only retained existing EXP2 mixtures and excluded cross-partition mixtures as required to enforce zero exact source overlap

between the retained development pool and the diagnostic test set. No new WAV files, mixture combinations, labels, or mel-spectrogram features were generated for this analysis.

As shown in Table 7, enforcing source-note disjointness substantially reduces the retained test support and makes the task more difficult. Accuracy decreases from 93.73% to 77.87%, and macro-F1 decreases from 93.73% to 77.99%. This reduction is consistent with the original complete-instance-disjoint performance having benefited in part from the extensive source-note reuse quantified in Section 4. However, because enforcing source-note disjointness also reduced the retained training set from 59,210 to 32,456 mixtures, the observed performance difference cannot be attributed exclusively to source reuse. Under the stricter unseen-source setting, performance remains above 72% F1 for each of the four triad classes. Per-class F1 scores under the diagnostic protocol are 73.50% for major, 72.38% for minor, 85.47% for diminished, and 80.60% for augmented. These results do not prove or disprove shortcut learning. Rather, they show that EXP2 source-note-disjoint evaluation is substantially more challenging than the complete-instance-disjoint benchmark, while the retained diagnostic performance indicates that triad quality remains partly recoverable from unseen exact source-note recordings within the controlled SECD domain.

Table 6. EXP2 test results: triad chord quality recognition.

Class	Support	Precision (%)	Recall (%)	F1 (%)
major	3161	91.53	90.95	91.24
minor	3161	90.32	91.46	90.88
diminished	3159	96.53	95.98	96.25
augmented	3207	96.57	96.51	96.54
Macro avg	12,688	93.74	93.72	**93.73**
Accuracy				**93.73**

Table 7. Comparison of the complete-instance-disjoint EXP2 benchmark with the complementary source-note-disjoint diagnostic. Changes are reported in percentage points (pp).

Metric	Benchmark EXP2	Source-Note-Disjoint Diagnostic
Test samples	12,688	1148
Accuracy	93.73%	77.87% (−15.86 pp)
Macro-F1	93.73%	77.99% (−15.74 pp)

5.4. EXP3: Per-Instrument Dynamics Classification

The third task evaluates 5-class per-instrument dynamics recognition under an arco-normal-only condition. The target classes are pianissimo (pp), piano (p), mezzo-piano (mp), forte (f), and fortissimo (ff). These labels correspond to the most frequent static dynamic categories available within the arco-normal portion of SECD, which is the dominant playing-technique condition in the corpus. The experiment therefore focuses on the best-supported dynamic-level labels under a controlled technique setting.

The full SECD metadata contains eight dynamic labels: pianissimo, piano, mezzo-piano, mezzo-forte, forte, fortissimo, crescendo, and decrescendo. EXP3 does not attempt to cover the complete dynamic vocabulary. Instead, it defines a focused benchmark over the five most populated static dynamic classes within the arco-normal subset. This choice reduces sparsity, avoids unstable estimates for underrepresented labels, and keeps the task centred on dynamic-level recognition rather than on modelling all dynamic markings in the full corpus.

Restricting the evaluation to arco-normal samples reduces technique-driven variability and focuses the task on the extent to which dynamic level can be identified from the

audio representation. Results are reported in a pooled voice-level setting across all present instrument voices.

This isolation is only partial. Perceptual and acoustic studies of timbre indicate that dynamic level can itself alter spectral structure, including brightness-related descriptors; therefore, the task should not be interpreted as simple amplitude classification [4]. Although the arco-normal condition removes one major source of playing-technique variation, the remaining dynamic categories may still be encoded through a combination of energy, spectral distribution, and instrument-specific response.

Table 8 reports the held-out test performance. The model reaches 98.19% accuracy and 98.01% macro-F1 over 71,898 pooled voice-level observations. The strongest results are obtained for the louder dynamic levels, with *f* and *ff* reaching F1 scores of 99.27% and 99.56%, respectively. The softer and intermediate categories also remain highly accurate, but show slightly greater confusion, especially between adjacent dynamic levels.

Table 8. EXP3 test results: per-instrument dynamics classification under the arco-normal-only condition, pooled across present instrument voices.

Class	Support	Precision (%)	Recall (%)	F1 (%)
pp	16,296	97.42	97.26	97.34
p	9829	95.96	96.08	96.02
mp	13,511	97.97	97.76	97.86
f	14,821	99.27	99.28	99.27
ff	17,441	99.44	99.67	99.56
Macro avg	71,898	**98.01**	**98.01**	**98.01**
Accuracy				**98.19**

Performance is not uniform across instruments. Cello achieves 99.72% macro-F1, viola 98.75%, and violin 1 97.45%, whereas violin 2 reaches 87.10% macro-F1. Violin 2 appears only in quartet-like four-voice textures and contributes 4,059 test observations, compared with 22,613 observations for each of cello, viola, and violin 1. The present evaluation does not isolate the relative contributions of the fixed violin 2 transformation, its smaller sample support, quartet-only context, instrument role, and overall mixture configuration. Within violin 2, the main difficulty concerns adjacent softer categories, especially *p* and *mp*, which obtain F1 scores of 76.83% and 79.68%, respectively.

Figure 5 shows the normalised pooled confusion matrix. The matrix confirms that most errors occur between adjacent dynamic categories. Confusion is concentrated around the softer and intermediate levels, while *f* and *ff* are almost perfectly separated in the pooled test evaluation.

5.5. EXP4: Playing Technique-Family Recognition

The fourth task evaluates 4-class playing-technique-family recognition. Each present instrument voice is assigned one of four family labels: bowed, col legno, harmonic, or plucked. These labels are not stored directly in the underlying dataset; they are derived deterministically from the filename-encoded atomic playing-technique annotations described in Section 3. EXP4 is therefore a classification problem over grouped metadata rather than over a native annotation layer of the SECD.

EXP4 does not attempt fine-grained 21-class technique recognition. Instead, all 21 filename-encoded atomic technique labels observed in the SECD metadata are mapped deterministically into four broader technique families. This family-level formulation is a methodological choice: the four target classes represent coarse sound-production regimes rather than exhaustive expressive categories. In string performance, bowing and articulation choices shape attack, spectrum, sustain, and perceived timbre, and thus provide

acoustically meaningful targets for audio classification [1,4,17]. At the same time, individual technique labels may vary substantially across instruments, registers, dynamics, and source-note realisations. Grouping the atomic labels into broader families therefore provides a more conservative and acoustically stable benchmark than treating every individual playing technique as an independent class.

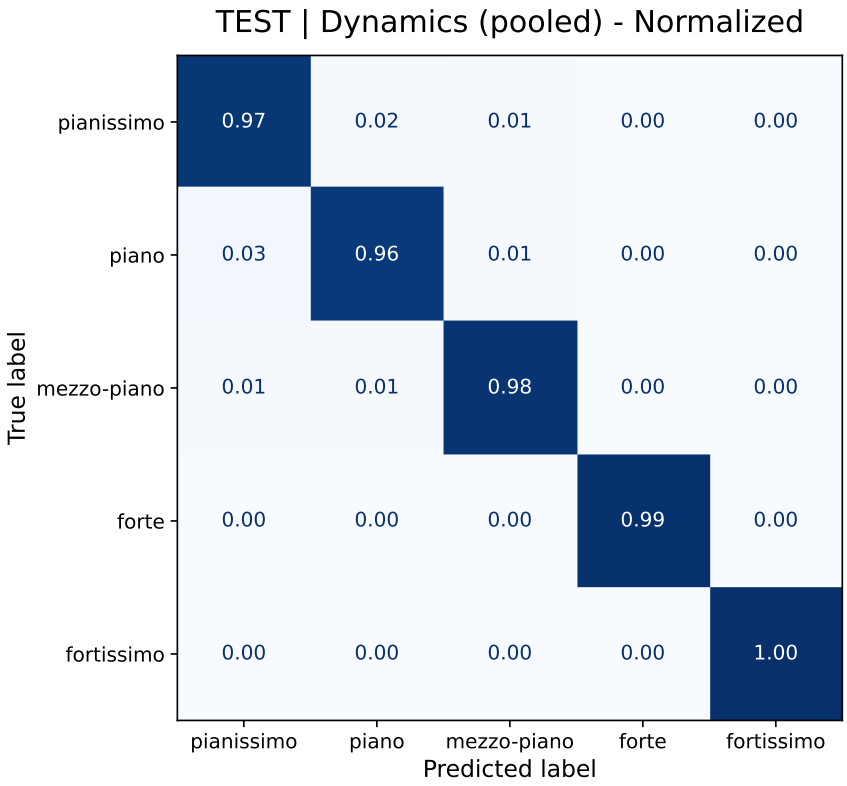

Figure 5. EXP3 normalised confusion matrix for pooled per-instrument dynamics classification on the held-out test set. Rows denote true labels and columns denote predicted labels. Darker cells indicate higher row-normalised values, while lighter cells indicate lower values.

The task is evaluated in a pooled voice-level setting across all present instrument voices. The held-out test set contains 129,121 labelled voice instances. The class distribution is strongly imbalanced: bowed samples constitute approximately 90% of the test set, with 116,273 instances, whereas col legno, harmonic, and plucked contain 3611, 3173, and 6064 instances, respectively.

This imbalance is an expected consequence of both musical practice and corpus construction, as introduced in Section 3. At the atomic-label level, `arco-normal` is the default sound-production mode for bowed string instruments. At the family level, `arco-normal` and other bow-mediated techniques are grouped into the broader bowed class. Because the professionally recorded solo-note pool contains substantially more standard bowed material than special-technique material, SECD inherits this source-level distribution through combinatorial generation and amplifies it at the instance and pooled voice-instance levels. Accordingly, EXP4 is framed as a family-level evaluation of broad sound-production regimes under realistic corpus imbalance, not as a balanced fine-grained technique-recognition benchmark.

The deterministic mapping used in EXP4 is shown in Table 9. The mapping covers all 21 filename-encoded atomic playing-technique labels represented in the EXP4 metadata after task-specific filtering.

Table 9. Deterministic mapping from the 21 filename-encoded atomic SECD playing-technique labels to the four technique families used in EXP4.

Family	Constituent Techniques
bowed	arco-normal, arco-sul-ponticello, arco-sul-tasto, arco-au-talon, arco-tremolo, arco-martele, arco-minor-trill, arco-major-trill, arco-glissando, molto-vibrato, non-vibrato, con-sord
col-legno	arco-col-legno-battuto, arco-col-legno-tratto
harmonic	natural-harmonic, artificial-harmonic, arco-harmonic
plucked	pizz-normal, pizz-glissando, pizz-tremolo, snap-pizz

Table 10 reports the held-out test performance. The model reaches 99.39% accuracy, but this value is strongly influenced by the bowed majority class. The macro-averaged F1 score of 97.29% therefore provides the more informative summary of performance across the four technique families. The bowed class is classified near-perfectly, with 99.68% F1. Col legno is the most challenging class, with 93.84% F1, while harmonic and plucked remain high at 97.79% and 97.85% F1, respectively.

Table 10. EXP4 test results: pooled voice-level playing-technique-family recognition.

Class	Support	Precision (%)	Recall (%)	F1 (%)
bowed	116,273	99.54	99.82	99.68
col_legno	3611	95.60	92.14	93.84
harmonic	3173	98.13	97.45	97.79
plucked	6064	99.25	96.49	97.85
Macro avg	129,121	98.13	96.47	**97.29**
Weighted avg	129,121	99.39	99.39	99.39
Accuracy				**99.39**

Figure 6 shows the row-normalised confusion matrix. The matrix confirms that the dominant bowed class is not merely inflating the pooled score: harmonic and plucked are also classified with high recall. The main residual error is the partial absorption of minority classes into the bowed category. Col legno shows the highest error rate, with approximately 8% of its instances predicted as bowed. Harmonic and plucked show smaller but similar error patterns, with approximately 3% of each class assigned to bowed.

Per-instrument results require careful interpretation because the per-instrument label space is not uniformly populated across all voices. Cello contains no plucked test instances, and viola contains no col legno test instances. As a result, per-instrument macro-F1 is not directly comparable across instruments, since some class-level scores are undefined for instruments with zero support in a given family. Weighted F1 and class-level recall are therefore more informative for these diagnostics. Under this interpretation, cello and viola remain strong on the classes present in their test partitions, with weighted F1 scores of 99.94% and 99.73%, respectively. Violin 2 remains the most difficult voice-level case, with 91.87% macro-F1 and 97.68% weighted F1, while violin 1 achieves 97.66% macro-F1 and 99.16% weighted F1.

Overall, EXP4 shows that broad technique-family recognition is highly reliable in the SECD, even under a strongly imbalanced class distribution inherited from both string-performance practice and the available source-note recordings. The remaining errors are concentrated in acoustically plausible confusions between minority sound-production regimes and bowed samples, rather than in systematic failures across all classes.

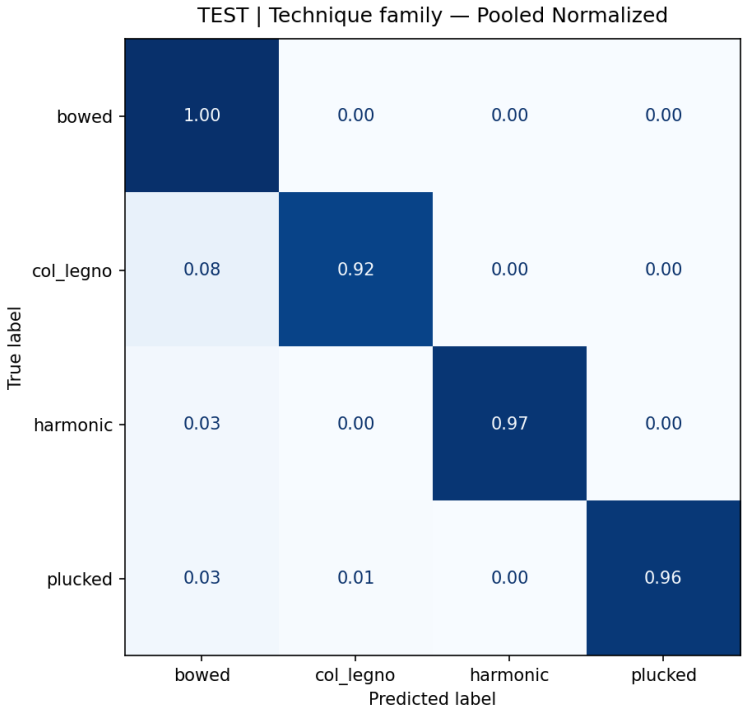

Figure 6. Row-normalised confusion matrix for EXP4 playing-technique-family recognition on the held-out test set. The main residual error is the absorption of minority technique families into the bowed category, especially for col legno. Darker cells indicate higher row-normalised values, while lighter cells indicate lower values.

5.6. Summary

Table 11 consolidates results across the four benchmark experiments. To our knowledge, these are the first reported reference results under this specific controlled formulation for each task. They should be interpreted as in-domain reference performance under the complete-instance-disjoint benchmark splitting procedure described in Section 4, rather than as upper bounds for source-note-disjoint or external-recording generalisation.

Because the four tasks differ substantially in class balance, Table 11 also reports a majority-class reference value for each experiment. This value is computed from the held-out test-set supports and corresponds to the accuracy obtained by always predicting the most frequent class. It is included as an interpretive reference rather than as a separately trained baseline model, and is especially important for EXP4, where the bowed family dominates the pooled voice-level distribution.

Table 11. Summary of benchmark results under the in-domain evaluation protocol. Majority reference denotes the accuracy obtained by always assigning every test instance to the most frequent class in that experiment's held-out test set; it is computed from the class supports and is not a separately trained model.

Exp	Task	Classes	Maj. Ref. (%)	Acc. (%)	Macro-F1 (%)
EXP1	Ensemble size	3	35.26	98.67	98.64
EXP2	Chord quality	4	25.28	93.73	93.73
EXP3	Dynamics (arco-normal only)	5	24.26	98.19	98.01
EXP4	Technique family	4	90.05	99.39	97.29

The AST-based models exceed the majority-class reference by a wide margin in all four experiments. For EXP1, EXP2, and EXP3, the reference values are low because the held-out test sets are relatively well distributed across classes. EXP4 requires more careful interpretation: assigning every test instance to the bowed family would already yield

90.05% accuracy. However, the model reaches 99.39% accuracy and 97.29% macro-F1, indicating that performance is not explained solely by the dominant bowed class and that the minority technique families are also recognised with high reliability.

6. Potential Applications and Research Use Cases

The SECD and its associated benchmarks support research directions in music information retrieval, audio representation learning, computational musicology, and music education technology. The use cases discussed below should be understood as research-oriented applications: they identify tasks for which the SECD provides controlled training, evaluation, or diagnostic material, rather than claiming immediate deployment readiness. In each case, the relevant contribution of the SECD lies in the combination of constructed polyphonic string mixtures, deterministic metadata, and per-voice annotations for pitch, dynamics, and playing technique.

6.1. Controlled Harmonic and Ensemble-Structure Analysis

The SECD can support controlled research on harmonic- and ensemble-structure analysis in bowed-string audio. Most audio chord-recognition work addresses continuous musical recordings, where chord labels must be estimated over time and where segmentation, key context, harmonic rhythm, and global musical structure affect the recognition problem [19–21]. The SECD defines a complementary setting: isolated harmonic intervals and chord events are rendered as polyphonic string mixtures with known ensemble size, chord or interval label, voicing condition, instrument assignment, and per-voice metadata. This makes it possible to study chord-quality and voice-multiplicity recognition without the additional confounds of full-song segmentation and long-range harmonic tracking.

This direction is represented by the EXP1 and EXP2 benchmarks. EXP1 evaluates whether the number of simultaneously sounding string voices is recoverable from a monaural mixture signal, while EXP2 evaluates whether triad quality remains recoverable when chord tones are realised by different string instruments across both loose and strict triad metadata regimes. The observed major–minor confusions in EXP2 are informative because these two categories differ by a single semitone in one chord voice. Such confusions indicate that the SECD captures a fine-grained acoustic discrimination problem in which symbolic harmonic differences are embedded within overlapping string spectra. Beyond the representative EXP1 and EXP2 benchmarks, the released ontology and metadata support future work on harmonic interval recognition, seventh-chord classification, and inversion-aware recognition. These tasks extend the benchmark space enabled by the SECD but are not evaluated in the present dataset-introduction study.

The same material can also support computational musicology and acoustic-perception studies. Because the harmonic labels, instrument roles, and metadata fields are controlled by construction, confusion patterns can be analysed as corpus-internal evidence about which chord or interval categories become acoustically closer under a given string-ensemble construction regime. This does not replace human listening experiments, but it can generate testable hypotheses about the relationship between symbolic harmonic categories and realised string sound. Such a framing is consistent with perception research showing that interval and chord perception depend not only on abstract pitch distance, but also on register, timbre, grouping, harmonicity, roughness, and listener context [4,25,26].

6.2. Per-Voice Expressive Attribute Recognition

A second use case concerns recognition of expressive per-voice attributes in polyphonic string mixtures. The SECD is not limited to chord labels: each present instrument voice is also associated with dynamic and playing-technique metadata. This makes it possible

to study whether models can recover attributes that belong to individual voices even when the input is a single mixed waveform. EXP3 addresses this problem through per-instrument dynamics classification under an arco-normal-only condition, while EXP4 addresses technique-family recognition by mapping the 21 atomic technique labels into four broader sound-production families.

This direction is relevant because dynamics and playing techniques are not merely symbolic annotations. Changes in dynamic level can alter both signal energy and spectral structure, while bowing, pizzicato, harmonics, col legno, vibrato-related categories, and mute-related conditions shape the temporal envelope, spectral distribution, and perceived timbre of string sound [1,4,17]. The SECD therefore enables controlled evaluation of whether such performance-related attributes remain detectable when several string voices are combined into a single acoustic scene.

This use case also has implications for music education technology and performance-feedback research. The SECD does not model live ensemble rehearsal interaction, but it provides a controlled intermediate setting in which per-instrument dynamics and broad technique-family labels are available inside polyphonic mixtures. Future educational systems could use SECD-like material for pre-training or diagnostic evaluation before being tested on independently recorded student or ensemble data. Such deployment-oriented claims require external validation and should not be inferred directly from the present in-domain benchmark results.

6.3. Multi-Task Representation Learning and Synthetic-to-Real Transfer

The SECD can serve as a specialised multi-task benchmark for evaluating audio representations on closely related string-ensemble tasks. MARBLE formalises the need for broad and reproducible evaluation of music audio representations, defining a taxonomy that spans acoustic, performance, score-level, and high-level MIR tasks [46]. The SECD is narrower in scope, but deeper within one controlled domain: the same constructed string-mixture representation can be evaluated for ensemble size, triad quality, per-instrument dynamics, and playing-technique family.

This structure enables controlled experiments on representation sharing. A single backbone can be tested for sensitivity to harmonic structure, voice multiplicity, dynamic level, and sound-production family. Such experiments can ask whether these attributes are mutually reinforcing, independent, or interfering when learned jointly from the same audio input. This is particularly relevant for transformer-based audio models, where a shared representation may encode multiple musical dimensions at once. Beyond the independent instrument-specific heads used in the present reference baselines, future work can investigate structured multi-voice models that explicitly represent dependencies among simultaneously sounding instrumental voices.

The SECD is also relevant to synthetic-to-real transfer research. Its mixtures are synthetic in the sense that they are generated by linear superposition, but they are constructed from professionally recorded instrumental notes rather than from purely symbolic or MIDI-rendered synthesis. This places the SECD between isolated-note corpora and real ensemble recordings: it offers exact label control while retaining many acoustic properties of recorded string tones. Recent work has shown that artificial polyphonic mixtures created from monophonic recordings can be effective for pre-training polyphonic transcription models [47], while MT3 demonstrates the relevance of transformer-based transfer learning for multi-instrument transcription under low-resource conditions [48]. Future work can therefore evaluate whether SECD-trained models transfer to independently recorded string ensemble data and extend source-note-disjoint analysis beyond EXP2 to the remaining benchmark tasks.

6.4. Attribute-Aware Source Attribution and Separation Research

The SECD may also support research on source attribution and attribute-aware separation, although it should not be presented as a conventional source-separation dataset unless separated stems or reconstruction targets are explicitly released. Its immediate contribution is different: each constructed mixture is associated with the instrumental role, pitch, dynamic marking, and playing technique of every present voice. This enables auxiliary supervision for models that attempt to infer which instrument contributes which musical or expressive attribute within a mixture.

This direction is complementary to synthetic chamber-ensemble separation datasets such as EnsembleSet, which was introduced to address the limited availability of sizeable, bleed-free multitrack datasets for separating similar-sounding ensemble sources [49]. EnsembleSet emphasises continuous chamber-ensemble material, multitrack rendering, and source separation. The SECD instead emphasises isolated harmonic events with explicit per-voice symbolic and performance-related metadata. The two dataset paradigms therefore address adjacent but distinct needs: one is oriented toward separation of continuous ensemble recordings, while the other enables controlled analysis of harmonic, instrumental, dynamic, and technique attributes inside short polyphonic string events.

Future work could examine whether auxiliary prediction of ensemble size, instrument participation, dynamics, or technique family improves the internal representations used by separation or source-attribution systems. This frames the SECD as a metadata-rich diagnostic corpus for attribute-aware mixture analysis rather than as a direct replacement for multitrack separation benchmarks.

6.5. Reproducible Benchmarking and Dataset Diagnostics

Finally, the SECD provides a reproducible testbed for controlled MIR benchmarking. Because the dataset is generated through explicit construction rules and exposes structured metadata for each mixture, researchers can design ablation studies that isolate the effects of ensemble size, harmonic type, strict versus loose metadata consistency, dynamic level, playing technique, and source-note reuse. This is useful for diagnosing whether a model is learning musically meaningful structure or exploiting easier corpus-specific regularities.

The distinction between the primary SECD corpus, the experimental mel-spectrogram cache, and the mini-SECD demonstration package is central to this use case. The full corpus supports acoustic experiments on constructed WAV mixtures and metadata. The feature cache supports efficient reproduction of the AST-based experiments. The mini-SECD package verifies the code path, data loading, splitting, collation, and model execution without requiring the full dataset download. Together, these resources allow researchers to reproduce the benchmark pipeline, test alternative models, and inspect how changes to the evaluation protocol affect reported performance.

This diagnostic role is especially important given the controlled source-note reuse discussed in Section 4. Because the SECD preserves explicit constituent-source references together with structured per-voice metadata, researchers can construct reproducible evaluation partitions that address different, explicitly bounded evaluation questions within the same underlying corpus. Complete-instance-disjoint partitions evaluate previously unseen mixtures assembled from the controlled source pool, whereas source-note-disjoint partitions evaluate mixtures composed entirely of exact source-note recordings not observed during model development. The appropriate protocol depends on the research question, because each evaluates a different and explicitly bounded form of generalisation. The EXP2 source-note-disjoint diagnostic presented in this work illustrates how the effect of this protocol choice can be quantified empirically. The SECD therefore contributes not

only benchmark tasks and reference results, but also the metadata required for transparent, reproducible, and research-question-specific evaluation design.

7. Discussion

The strong complete-instance-disjoint, in-domain performance of the AST-based baselines across all four tasks indicates that pre-trained transformer audio representations capture discriminative spectro-temporal information relevant to the selected SECD classification tasks, even though AudioSet contains no annotations specifically targeting classical string instrument attributes. This strong in-domain baseline performance is consistent with prior work on music tagging [34] and contrastive audio representation learning [35]. The EXP1–EXP4 benchmark results should be interpreted within this declared protocol and do not establish source-note-disjoint or independently recorded ensemble generalisation. The complementary EXP2 diagnostic provides a narrower evaluation of triad-quality recognition using unseen exact source-note recordings; its findings should not be extrapolated to the other benchmark tasks or to external ensemble recordings.

These results should be interpreted in relation to the specific ontology implemented by the SECD. The benchmark labels are musically meaningful, but they operate at a controlled and deliberately simplified level: chord quality is defined over rendered pitch combinations, dynamics over nominal markings, and technique over broad sound-production families. In live performance, these properties interact with register, phrasing, ensemble balance, articulation detail, and listener expectations [2,4,17]. The SECD is therefore best understood as a high-control benchmark for studying how such attributes are reflected in audio, not as a complete model of string quartet performance.

The four benchmark tasks probe different aspects of this controlled acoustic setting. EXP1 evaluates whether the model can infer voice multiplicity from spectral density and overlapping string sonorities. EXP2 evaluates whether harmonic quality remains recoverable when chord tones are realised by different string instruments across both loose and strict triad metadata regimes. EXP3 tests whether nominal dynamic markings remain encoded in the mixture beyond simple amplitude differences, while EXP4 evaluates whether broad sound-production regimes can be recovered despite severe family-level imbalance. Taken together, the tasks show that the SECD is not a single-purpose chord dataset, but a multi-attribute benchmark for studying harmonic, timbral, dynamic, and textural information in polyphonic string audio.

The strongest results occur in EXP1, EXP3, and EXP4, whereas EXP2 is comparatively more difficult. This ordering is musically plausible. Ensemble size, dynamic level, and broad technique family can produce distributed spectro-temporal signatures, while triad quality often depends on smaller pitch-class differences embedded within overlapping harmonic spectra. The major–minor confusions observed in EXP2 are therefore informative: they indicate that the benchmark captures a fine-grained acoustic discrimination problem rather than merely a coarse chord-labelling task.

The EXP4 results also illustrate why accuracy alone is insufficient for the SECD. Because the bowed family dominates the pooled voice-level distribution, the high overall accuracy must be interpreted together with macro-F1 and the class-normalised confusion matrix. The strong minority-class results suggest that the model is not simply exploiting the bowed majority. However, the residual tendency of minority classes to be absorbed into the bowed category shows that family-level imbalance remains a modelling challenge.

Several limitations qualify these conclusions. First, the SECD is constructed by linear superposition of pre-recorded isolated notes and therefore does not include the interactional mechanisms of live string ensemble performance. It excludes performer feedback, adaptive intonation, phrase-level timing negotiation, gesture-mediated coordination, room

interaction, sympathetic resonance, and changing bowing decisions that arise when musicians respond to one another in real time. Work on string quartet interdependence shows that real ensemble performance can be characterised through coupled behaviour across timing, intonation, dynamics, timbre, and articulation [3]. The SECD should therefore be interpreted as a controlled acoustic corpus for chord-level string-mixture analysis. Generalisation from SECD-trained models to independently recorded ensemble performance must be evaluated empirically.

Second, the compositional construction of the SECD means that individual source-note recordings may reappear across multiple interval and chord instances, including across the benchmark partitions. This is a structural property of a controlled compositional corpus generated from a finite source pool and should be distinguished from duplicate-instance leakage. The EXP2 audit quantified extensive source reuse in the original split, while the complementary diagnostic enforced zero exact source overlap between the retained development pool and the diagnostic test set after excluding cross-partition mixtures. Under this stricter protocol, accuracy decreased from 93.73% to 77.87%, and macro-F1 decreased from 93.73% to 77.99%. This reduction indicates that the reported EXP2 performance likely benefited in part from source reuse. However, because enforcing source-note disjointness also substantially reduced the retained training set, the observed performance decrease cannot be attributed exclusively to source reuse. The retained diagnostic performance nevertheless indicates that triad quality remains partly recoverable from unseen exact source-note recordings within the controlled SECD domain. Because the diagnostic is restricted to EXP2 and uses a substantially smaller retained subset, it should be interpreted as targeted supplementary evidence rather than as source-note-disjoint validation of all SECD benchmark tasks.

Third, the distribution of playing techniques is dominated by standard bowed material, especially `arco-normal`. This reflects the dual origin of the corpus: normal bowing is the principal sound-production mode of bowed string performance, and the available professionally recorded source-note pool contains more standard bowed examples than rare or specialised techniques. The EXP4 results should therefore be read as performance under realistic family-level imbalance, with macro-F1 reported alongside accuracy to avoid overstating performance on the bowed majority class.

Fourth, the violin 2 result in EXP3 (87.10% macro-F1) highlights the importance of considering per-instrument support and role-specific construction when interpreting multi-head classification results. Violin 2 has substantially lower support and shows greater difficulty on adjacent dynamic levels, but the present evaluation does not isolate the relative contributions of sample support, the fixed violin 2 transformation, quartet-only context, instrument role, or overall mixture configuration. Data augmentation strategies [50] may be relevant to this setting, but their effect remains to be evaluated in future work.

Finally, the full SECD corpus provides substantial material for additional tasks beyond the four representative benchmarks reported here, including seventh-chord classification and harmonic interval recognition, with 101,278 and 101,224 available samples, respectively. These tasks, together with inversion-aware recognition, constitute natural directions for extending the SECD benchmark suite in future work.

These limitations also point to the SECD's broader methodological value. Because the dataset construction process is explicit and controllable, the corpus can support staged experiments on label granularity, source-note overlap, synthetic-to-real transfer, and the trade-off between ecological validity and annotation precision. In that sense, the SECD contributes not only benchmark numbers but also a reproducible framework for asking which musical attributes remain identifiable once multiple string voices are combined into a single acoustic scene.

8. Reproducibility and Data Availability

Throughout this paper, we distinguish between the primary SECD corpus, the experimental feature cache, and the mini-SECD demonstration package. The primary SECD corpus consists of constructed WAV mixture files and accompanying metadata and is released through Zenodo. The experimental feature cache consists of mel spectrogram tensors derived from the SECD WAV mixtures for the AST-based experiments; these tensors are implementation artifacts of the benchmark pipeline rather than the primary dataset format. The mini-SECD package is provided through the project GitHub repository and includes metadata-compatible empty WAV placeholders and pre-computed mel spectrograms, allowing the codebase to be executed without downloading or processing the full SECD WAV corpus.

8.1. Mini-SECD Demonstration Package

To support code-level reproducibility without requiring the full SECD corpus, we provide mini-SECD, a compact demonstration package containing approximately 1600 pre-computed mel spectrogram tensors balanced across the four benchmark settings. Mini-SECD includes reconstructed CSV metadata files, metadata-compatible empty WAV placeholders with valid audio headers, and a folder structure mirroring the full dataset. The placeholder WAV files preserve path compatibility in the codebase; the acoustic inputs used by the demonstration pipeline are the included pre-computed mel spectrograms. Mini-SECD therefore verifies the benchmark code path, including data loading, splitting, collation, and model execution, but is not intended as an acoustically complete subset of the full SECD WAV corpus.

8.2. Split Generation

Experimental train/validation/test splits are generated within each benchmark pipeline after task-specific filtering and label mapping. The resulting split indices are stored to disk and reused for exact reproduction of the reported EXP1–EXP4 runs.

The complementary EXP2 source-note-disjoint diagnostic used a separate source-controlled partition constructed from the existing EXP2 mixtures according to the eligibility and exclusion rules described in Section 4. The same EXP2 model architecture, preprocessing, cached features, label definition, and training configuration were then applied to the retained train/validation/test partitions.

8.3. Licence and Access

The full SECD corpus is publicly released through Zenodo and consists of constructed WAV mixture files and accompanying metadata. The original isolated Philharmonia Orchestra solo recordings are not redistributed as part of the SECD. The SECD should therefore be understood as a derived mixture corpus constructed from the source material, not as a redistribution of the original solo-note library.

The project repository provides split-generation code, benchmark scripts, and the mini-SECD demonstration package. The conditions of reuse for repository materials are specified in the repository licence. Metadata and other non-audio derivative files are released under the licence specified in the public repository. The licensing status of the constructed SECD WAV mixtures is stated explicitly in the Zenodo record.

9. Conclusions

We have presented the SECD, a large-scale annotated dataset comprising 287,088 string ensemble harmonic-interval and chord instances for controlled MIR benchmarking. The dataset supports four representative reference classification tasks: ensemble size recog-

nition, chord quality identification, per-instrument dynamics classification, and playing technique-family recognition. AST-based baseline models achieve strong complete-instance-disjoint, in-domain performance across these tasks, with accuracies ranging from 93.73% to 99.39%, establishing, to our knowledge, the first reported results under this specific controlled benchmark formulation.

The paper also provides an explicit methodological account of controlled source-note reuse in datasets constructed from finite pools of isolated recordings. For EXP2, we quantified the extensive source overlap in the original complete-instance-disjoint split and performed a complementary source-note-disjoint diagnostic with zero exact source overlap between the retained development pool and the diagnostic test set. Performance decreased substantially under this stricter evaluation, indicating that the reported EXP2 result likely benefited in part from source reuse, while triad quality remained partly recoverable from unseen exact source-note recordings. This diagnostic constitutes targeted supplementary evidence for EXP2 and should not be interpreted as source-note-disjoint validation of the other benchmark tasks or as evidence of generalisation to independently recorded ensemble audio.

More broadly, the constituent-source information and structured per-voice metadata preserved by the SECD enable researchers to construct reproducible evaluation partitions aligned with different, explicitly bounded research questions. Future work can extend source-note-disjoint evaluation to the other SECD tasks, expand the benchmark suite to seventh-chord classification, harmonic-interval recognition, and inversion-aware chord recognition, investigate structured multi-voice modelling, and validate SECD-trained models on independently recorded ensemble audio. The SECD therefore provides a controlled and reproducible research platform for fine-grained string-ensemble audio analysis and for systematically examining how evaluation design affects measured performance.

Author Contributions: Conceptualization, A.G.; methodology, A.G., P.Z. and G.T.; software, A.G.; validation, A.G., P.Z. and G.T.; formal analysis, A.G.; investigation, A.G.; resources, A.G.; data curation, A.G.; writing—original draft preparation, A.G.; writing—review and editing, A.G., P.Z. and G.T.; visualization, A.G.; supervision, P.Z. and G.T.; project administration, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full SECD corpus is available through Zenodo as constructed WAV mixture files with accompanying metadata at https://doi.org/10.5281/zenodo.15547207. The original isolated Philharmonia Orchestra solo recordings are not redistributed as part of the SECD. The project repository is available at https://github.com/ageroul/SECD (accessed on 1 July 2026) and provides split-generation code, benchmark scripts, saved split definitions for the reported EXP1–EXP4 runs, and the mini-SECD demonstration package for lightweight reproducibility. Users of the SECD should cite the Zenodo dataset record and this paper.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.5 Thinking) for English-language editing, including grammar checking and minor wording and clarity improvements. No generative artificial intelligence tools were used to generate experimental data, perform analyses, compute results, create figures, design the study, or draw scientific conclusions. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goodman, E. Ensemble Performance. In *Musical Performance: A Guide to Understanding*; Rink, J., Ed.; Cambridge University Press: Cambridge, UK, 2002; pp. 153–167. [CrossRef]
2. Palmer, C. Music Performance: Movement and Coordination. In *The Psychology of Music*, 3rd ed.; Deutsch, D., Ed.; Academic Press: Cambridge, MA, USA, 2013; pp. 405–422. [CrossRef]
3. Papiotis, P.; Marchini, M.; Pérez-Carrillo, A.; Maestre, E. Measuring Ensemble Interdependence in a String Quartet through Analysis of Multidimensional Performance Data. *Front. Psychol.* **2014**, *5*, 963. [CrossRef] [PubMed]
4. McAdams, S. Musical Timbre Perception. In *The Psychology of Music*, 3rd ed.; Deutsch, D., Ed.; Academic Press: Cambridge, MA, USA, 2013; pp. 35–67. [CrossRef]
5. Goto, M.; Hashiguchi, H.; Nishimura, T.; Oka, R. RWC Music Database: Popular, Classical, and Jazz Music Databases. In Proceedings of the 3rd International Society for Music Information Retrieval Conference, Paris, France, 13–17 October 2002; pp. 287–288.
6. Bittner, R.; Salamon, J.; Tierney, M.; Mauch, M.; Cannam, C.; Bello, J.P. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In Proceedings of the 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, 27–31 October 2014; pp. 155–160.
7. Bosch, J.J.; Fuhrmann, F.; Herrera, P. IRMAS: A Dataset for Instrument Recognition in Musical Audio Signals, Version 1.0. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012), Porto, Portugal*; Zenodo: Geneva, Switzerland, 2014. [CrossRef]
8. Li, B.; Liu, X.; Dinesh, K.; Duan, Z.; Sharma, G. Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Trans. Multimed.* **2019**, *21*, 522–535. [CrossRef]
9. Peeters, G. The Deep Learning Revolution in MIR: The Pros and Cons, the Needs and the Challenges. In *Perception, Representations, Image, Sound, Music*; Kronland-Martinet, R., Ystad, S., Aramaki, M., Eds.; Springer: Cham, Switzerland, 2021; Volume 12631, pp. 3–30. [CrossRef]
10. Lostanlen, V.; Andén, J.; Lagrange, M. Extended Playing Techniques: The Next Milestone in Musical Instrument Recognition. In Proceedings of the 5th International Conference on Digital Libraries for Musicology (DLfM '18), New York, NY, USA, 28 September 2018; pp. 1–10. [CrossRef]
11. Cella, C.E.; Ghisi, D.; Lostanlen, V.; Lévy, F.; Fineberg, J.; Maresz, Y. OrchideaSOL: A Dataset of Extended Instrumental Techniques for Computer-Aided Orchestration. *arXiv* **2020**, arXiv:2007.00763. [CrossRef]
12. Alex, T.; Atito, S.; Mustafa, A.; Awais, M.; Jackson, P.J.B. SSLAM: Enhancing Self-Supervised Models with Audio Mixtures for Polyphonic Soundscapes. In Proceedings of the The Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
13. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 571–575. [CrossRef]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
16. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780. [CrossRef]
17. Hill, P. From Score to Sound. In *Musical Performance: A Guide to Understanding*; Rink, J., Ed.; Cambridge University Press: Cambridge, UK, 2002; pp. 129–143.
18. Fonseca, E.; Favory, X.; Pons, J.; Font, F.; Serra, X. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 829–852. [CrossRef]
19. Korzeniowski, F.; Widmer, G. Feature Learning for Chord Recognition: The Deep Chroma Extractor. In Proceedings of the 17th International Society for Music Information Retrieval Conference, New York, NY, USA, 7–11 August 2016; pp. 37–43.
20. Korzeniowski, F.; Widmer, G. Improved Chord Recognition by Combining Duration and Harmonic Language Models. In Proceedings of the 19th International Society for Music Information Retrieval Conference, Paris, France, 23–27 September 2018; pp. 663–670.
21. McVicar, M.; Santos-Rodriguez, R.; Ni, Y.; De Bie, T. Automatic Chord Estimation from Audio: A Review of the State of the Art. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 556–575. [CrossRef]
22. Akram, M.W.; Dettori, S.; Colla, V.; Buttazzo, G.C. ChordFormer: A Conformer-Based Architecture for Large-Vocabulary Audio Chord Recognition. *arXiv* **2025**, arXiv:2502.11840. [CrossRef]

23. Park, J.; Choi, K.; Jeon, S.; Kim, D.; Park, J. A Bi-Directional Transformer for Musical Chord Recognition. In Proceedings of the 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 4–8 November 2019; pp. 620–627.
24. Grollmisch, S.; Cano, E.; Mora Ángel, F.; López Gil, G. Ensemble Size Classification in Colombian Andean String Music Recordings. In *Perception, Representations, Image, Sound, Music*; Kronland-Martinet, R., Ystad, S., Aramaki, M., Eds.; Springer: Cham, Switzerland, 2021; Volume 12631, pp. 60–74. [CrossRef]
25. Deutsch, D. (Ed.) The Processing of Pitch Combinations. In *The Psychology of Music*, 3rd ed.; Academic Press: Cambridge, MA, USA, 2013; pp. 249–325. [CrossRef]
26. Thompson, W.F. Intervals and Scales. In *The Psychology of Music*, 3rd ed.; Deutsch, D., Ed.; Academic Press: Cambridge, MA, USA, 2013; pp. 107–140. [CrossRef]
27. Han, Y.; Kim, J.; Lee, K. Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 208–221. [CrossRef]
28. Eggink, J.; Brown, G.J. Instrument Recognition in Accompanied Sonatas and Concertos. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, QC, Canada, 17–21 May 2004; Volume 5, pp. iv-217–iv-220. [CrossRef]
29. Choi, K.; Fazekas, G.; Sandler, M. Convolutional Recurrent Neural Networks for Music Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2011–2015. [CrossRef]
30. Li, D.; Ma, Y.; Wei, W.; Kong, Q.; Wu, Y.; Che, M.; Xia, F.; Benetos, E.; Li, W. MERTech: Instrument Playing Technique Detection Using Self-Supervised Pretrained Model with Multi-Task Finetuning. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing, Denver, CO, USA, 9–13 June 2024; pp. 521–525. [CrossRef]
31. Dalmazzo, D.; Ramírez, R. Bowing Gestures Classification in Violin Performance: A Machine Learning Approach. *Front. Psychol.* **2019**, *10*, 344. [CrossRef] [PubMed]
32. Papiotis, P.; Maestre, E.; Marchini, M.; Pérez-Carrillo, A. *QUARTET Dataset*; Zenodo: Geneva, Switzerland, 2021. [CrossRef]
33. Pons, J.; Nieto, O.; Prockup, M.; Schmidt, E.M.; Ehmann, A.F.; Serra, X. End-to-End Learning for Music Audio Tagging at Scale. In Proceedings of the 19th International Society for Music Information Retrieval Conference, Paris, France, 23–27 September 2018; pp. 637–644.
34. Won, M.; Ferraro, A.; Bogdanov, D.; Serra, X. Evaluation of CNN-Based Automatic Music Tagging Models. In Proceedings of the 17th Sound and Music Computing Conference, Turin, Italy, 24–26 June 2020.
35. Spijkervet, J.; Burgoyne, J.A. Contrastive Learning of Musical Representations. In Proceedings of the 22nd International Society for Music Information Retrieval Conference, Online, 7–12 November 2021; pp. 673–681.
36. Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Xiao, C.; Lin, C.; Ragni, A.; Benetos, E.; et al. MERT: Acoustic Music Understanding Model with Large-Scale Self-Supervised Training. In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
37. Engel, J.; Resnick, C.; Roberts, A.; Dieleman, S.; Norouzi, M.; Eck, D.; Simonyan, K. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1068–1077.
38. Duan, Z.; Pardo, B.; Zhang, C. Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-Peak Regions. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 2121–2133. [CrossRef]
39. Philharmonia Orchestra. Sound Samples. Online Resource. Available online: https://philharmonia.co.uk/resources/sound-samples/ (accessed on 10 March 2025).
40. Woodhouse, J.; Galluzzo, P.M. The Bowed String as We Know It Today. *Acta Acust. United Acust.* **2004**, *90*, 579–589.
41. Guettler, K. Bows, Strings, and Bowing. In *The Science of String Instruments*; Rossing, T.D., Ed.; Springer: New York, NY, USA, 2010; pp. 279–300.
42. Stowell, R. (Ed.) Extending the Technical and Expressive Frontiers. In *The Cambridge Companion to the String Quartet*; Cambridge University Press: Cambridge, UK, 2011; pp. 149–174.
43. Piston, W. *Harmony*; W. W. Norton & Company: New York, NY, USA, 1987.
44. Adler, S. *The Study of Orchestration*; W. W. Norton & Company: New York, NY, USA, 2002.
45. Weck, B.; Serra, X. Data Leakage in Cross-Modal Retrieval Training: A Case Study. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.
46. Yuan, R.; Ma, Y.; Li, Y.; Zhang, G.; Chen, X.; Yin, H.; Zhuo, L.; Liu, Y.; Huang, J.; Tian, Z.; et al. MARBLE: Music Audio Representation Benchmark for Universal Evaluation. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
47. Simon, I.; Gardner, J.; Hawthorne, C.; Manilow, E.; Engel, J. Scaling Polyphonic Transcription with Mixtures of Monophonic Transcriptions. In Proceedings of the 23rd International Society for Music Information Retrieval Conference, Bengaluru, India, 4–8 December 2022; pp. 44–51.

48. Gardner, J.; Simon, I.; Manilow, E.; Hawthorne, C.; Engel, J. MT3: Multi-Task Multitrack Music Transcription. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
49. Sarkar, S.; Benetos, E.; Sandler, M. EnsembleSet: A New High Quality Synthesised Dataset for Chamber Ensemble Separation. In Proceedings of the 23rd International Society for Music Information Retrieval Conference, Bengaluru, India, 4–8 December 2022; pp. 625–632.
50. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2613–2617. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.